UNIVERSIDAD NACIONAL AMAZÓNICA DE MADRE DE DIOS
FACULTAD DE INGENIERÍA
ESCUELA PROFESIONAL DE INGENIERÍA DE SISTEMAS E INFORMÁTICA

TESIS

"PREDICCIÓN DEL RENDIMIENTO ACADÉMICO EMPLEANDO ALGORITMOS DE APRENDIZAJE SUPERVISADO EN ESTUDIANTES DEL PRIMER SEMESTRE DE LA CARRERA PROFESIONAL DE INGENIERÍA DE SISTEMAS E INFORMÁTICA DE LA UNAMAD, 2020"

PARA OPTAR EL TÍTULO PROFESIONAL DE INGENIERO EN SISTEMAS E INFORMÁTICA

AUTOR:
Bach. VARGAS QUISPE, Alex Ali
ASESOR:
M.Sc. PRIETO LUNA, Jaime Cesar

Puerto Maldonado, noviembre 2022

UNIVERSIDAD NACIONAL AMAZÓNICA DE MADRE DE DIOS

FACULTAD DE INGENIERÍA

ESCUELA PROFESIONAL DE INGENIERÍA DE SISTEMAS E INFORMÁTICA

TESIS

"PREDICCIÓN DEL RENDIMIENTO ACADÉMICO EMPLEANDO ALGORITMOS DE APRENDIZAJE SUPERVISADO EN ESTUDIANTES DEL PRIMER SEMESTRE DE LA CARRERA PROFESIONAL DE INGENIERÍA DE SISTEMAS E INFORMÁTICA DE LA UNAMAD, 2020"

PARA OPTAR EL TÍTULO PROFESIONAL DE INGENIERO EN SISTEMAS E INFORMÁTICA

AUTOR:

Bach. VARGAS QUISPE, Alex Ali

ASESOR:

M.Sc. PRIETO LUNA, Jaime Cesar

Puerto Maldonado, noviembre 2022

DEDICATORIA

Le dedico esta tesis a mis seres queridos y a la Universidad Nacional Amazónica de Madre de Dios por ser parte importante en mi formación profesional.

AGRADECIMIENTO

Agradezco primero a dios por darnos salud y bienestar para poder realizar sin complicaciones esta tesis. De igual forma a mi familia y a mis amigos de la unidad del equipo de desarrollo de TI quienes fueron el pilar de apoyo en todo momento.

PRESENTACIÓN

El presente trabajo de investigación titulado "**Predicción del rendimiento académico empleando algoritmos de aprendizaje supervisado en estudiantes del primer semestre de la carrera profesional de Ingeniería de Sistemas e Informática de la UNAMAD, 2020**", para obtener el título en Ingeniería de Sistemas e Informática. La tesis tiene como objetivo determinar los factores que tienen mayor relación con el rendimiento académico de los estudiantes de la carrera profesional de Ingeniería de Sistemas e Informática de la Universidad Nacional Amazónica de Madre de Dios y emplear los algoritmos de aprendizaje supervisado para predecir el rendimiento académico. En ese escenario se procesó un conjunto de datos conformado por 861 registros de los estudiantes de la carrera profesional de Ingeniería de Sistemas e Informática. El resultado obtenido de esta investigación pretende apoyar en la toma de decisiones por parte de los responsables de las áreas académicas y administrativas.

Bach. VARGAS QUISPE, Alex Ali

RESUMEN

La presente investigación tiene como objetivo predecir el rendimiento académico de los estudiantes del primer semestre de la carrera profesional de Ingeniería de Sistemas e Informática mediante el uso de algoritmos supervisados. La investigación se desarrolló bajo un enfoque cuantitativo, con un diseño no experimental de tipo correlacional transversal. La muestra estuvo constituida por 861 registros reales de los estudiantes ingresantes a la carrera profesional de Ingeniería de Sistemas e Informática de la Universidad Nacional Amazónica de Madre de Dios recopilados durante los semestres académicos 2010-1 al 2020-2. Se emplearon tres modelos de aprendizaje automático: árbol de decisiones, K-NN y Naive Bayes para establecer la relación o asociación de condiciones relacionadas con el rendimiento académico lo que servirá como información complementaria sobre el proceso de aprendizaje de los estudiantes. Los resultados permitieron identificar que K-NN logró el mayor accuracy con 81.97%. Se concluye que las dimensiones sociales, económicas y académicas son los que más influyen en el rendimiento académico.

Palabras claves: aprendizaje supervisado, predicción, rendimiento académico

ABSTRACT

The objective of this research is to predict the academic performance of the students of the first semester of the professional career of Systems Engineering and Informatics through the use of supervised algorithms. The research was developed under a quantitative approach, with a non-experimental design of a cross-correlational type. The sample consisted of 861 real records of students entering the professional career of Systems Engineering and Informatics at the National Amazon University of Madre de Dios collected during the academic semesters 2010-1 to 2020-2. Three machine learning models were used: decision tree, K-NN and Naive Bayes to establish the relationship or association of conditions related to academic performance, which will serve as complementary information on the learning process of students. The results allowed us to identify that K-NN achieved the highest accuracy with 81.97%. It is concluded that the social, economic and academic dimensions are the ones that most influence academic performance.

Keywords: supervised learning, prediction, academic performance

INTRODUCCIÓN

Las instituciones de educación superior han ingresado a la era de la “big data" y están recopilando grandes volúmenes de datos relacionados con sus alumnos y las dimensiones educativas (Raza Hasan, 2018). La detección de las dimensiones educativas es de gran interés en entornos académicos, puesto que permite comprender las actividades de aprendizaje de los estudiantes, así como también mejorar los resultados de aprendizaje (Czibula et al., 2019).

Uno de los mayores desafíos es incrementar la calidad de los procesos educativos para mejorar el rendimiento de los estudiantes. Los resultados de la predicción pueden ayudar a los estudiantes a desarrollar una buena comprensión de su nivel de desempeño en un curso y tomar las medidas correspondientes. Los instructores pueden actualizar su metodología de enseñanza para cumplir con los requisitos de los estudiantes con bajo rendimiento y brindar orientación adicional (Imran et al., 2019).

Por otra parte, un objetivo a largo plazo que se proponen las instituciones educativas en todo el mundo es aumentar la retención de estudiantes, puesto que existen muchos impactos positivos al lograr una mayor retención, como incrementar la reputación y clasificación de la universidad, y mejores oportunidades laborales para los alumnos (Czibula et al., 2019).

En el contexto de la pandemia por coronavirus la Universidad Nacional Amazónica de Madre de Dios planteó algunas adecuaciones en el plan curricular para el desarrollo del semestre académico en la modalidad no presencial, en ese escenario virtual conocer el posible rendimiento académico de los estudiantes resultaba clave para los docentes y personal de gestión administrativa de la universidad para una oportuna toma de decisiones ante posibles problemas en el desempeño académico de los estudiantes.

En virtud de lo referido, se pretende emplear algoritmos de aprendizaje supervisado para predecir el rendimiento académico y determinar las dimensiones que tienen una relación con el rendimiento académico de los estudiantes de la carrera profesional de Ingeniería de Sistemas e Informática.

El presente trabajo de investigación se encuentra dividido en cuatro capítulos. En el primer capítulo se expone el problema de investigación, que incluye la formulación, objetivos, variables y su operacionalización, hipótesis, y la justificación. En el segundo capítulo se presenta el desarrollo del marco teórico de la investigación y se realiza una revisión del estado del arte, antecedentes nacionales e internacionales relacionados al tema de estudio para conocer los métodos y técnicas empleados en dichas investigaciones. En el tercer capítulo se plantea el marco metodológico bajo la cual se desarrolla el trabajo de campo de la variable de estudio en esta investigación, el diseño, población y muestra, las técnicas e instrumentos de recopilación de datos y el método de análisis. En el cuarto capítulo se interpretan los resultados obtenidos. Finalmente, se construyen las conclusiones, sugerencias y las referencias bibliográficas.

INDICE

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

CAPITULO I. PROBLEMA DE INVESTIGACIÓN

1.1. Descripción del problema

Las universidades a nivel mundial buscan implementar nuevos métodos y técnicas de enseñanza para mejorar el rendimiento académico y el nivel educativo (Abdallah et al., 2020). El rendimiento académico durante los últimos años se convirtió en una pieza fundamental en la educación superior, las instituciones universitarias están obligadas a afrontar retos importantes para darle solución en estos tiempos cambiantes, los desafíos pasan por reformular algunos planteamientos o tradiciones que las instituciones han adoptado a lo largo del tiempo (de Pablos Pons, 2018).

Por ello, es crítico determinar las dimensiones claves que influyen en el desempeño de los estudiantes en un entorno de enseñanza, con el objetivo de mejorar los resultados de aprendizaje de los estudiantes y lograr dotar a la sociedad de profesionales bien preparados y capaces de insertarse en el mercado laboral (Gonzalez-Nucamendi et al., 2021).

La predicción de la dimensión que tiene una mayor relación con el rendimiento académico es de gran relevancia práctica en los entornos educativos, ya que puede proporcionar una retroalimentación relevante para los estudiantes que probablemente reprobarán un determinado curso, y con una asesoría adicional durante el semestre estos estudiantes pueden prevenir un posible fracaso académico (Czibula et al., 2019).

El contexto de la pandemia por COVID-19 trajo consigo un cambio sin precedentes en el sistema educativo mundial, la generación de grandes volúmenes de datos (Abdallah et al., 2020). El uso de esta gran cuantía de datos para predecir el rendimiento de los estudiantes se convirtió en una tarea crucial (Imran et al., 2019).

La Universidad Nacional Amazónica de Madre de Dios no está ajena a estos cambios y ha logrado almacenar gran cantidad de datos de los estudiantes ingresantes en el transcurso de los años, esta información es muy importante

para su uso en diferentes campos. Sin embargo, esta información no es procesada ni analizada para dar soporte en la toma de decisiones de los técnicas de enseñanza, tutorías, incentivos o monitoreo del rendimiento de los estudiantes ingresantes.

1.2. Formulación del problema

1.2.1. Problema general

¿Cómo se puede predecir el rendimiento académico de los estudiantes del primer semestre de la carrera profesional de Ingeniería de Sistemas e Informática de la UNAMAD mediante algoritmos de aprendizaje supervisado?

1.2.2. Problemas específicos

1) ¿Cuáles son los indicadores sociales, económicos y académicos con mayor incidencia para predecir el rendimiento académico de los estudiantes del primer semestre de la carrera profesional de Ingeniería de Sistemas e Informática de la UNAMAD?
2) ¿Qué algoritmos de aprendizaje supervisado pueden predecir el rendimiento académico de los estudiantes del primer semestre de la carrera profesional de Ingeniería de Sistemas e Informática de la UNAMAD?

1.3. Objetivos

1.3.1. Objetivo general

Predecir el rendimiento académico de los estudiantes del primer semestre de la carrera profesional de Ingeniería de Sistemas e Informática de la UNAMAD mediante algoritmos de aprendizaje supervisado.

1.3.2. Objetivos específicos

1) Determinar los indicadores sociales, económicos y académicos con mayor incidencia para predecir el rendimiento académico de los estudiantes del

primer semestre de la carrera profesional de Ingeniería de Sistemas e Informática de la UNAMAD.

2) Determinar los algoritmos de aprendizaje supervisado que pueden predecir el rendimiento académico de los estudiantes del primer semestre de la carrera profesional de Ingeniería de Sistemas e Informática de la UNAMAD.

1.4. Variables

Variable independiente: Algoritmos de aprendizaje supervisado.

Variable dependiente: Rendimiento académico.

1.5. Operacionalización de variables

Tabla 1

Operacionalización de las variables

Variables	Definición conceptual	Dimensiones	Indicadores
Variable independiente: Algoritmos de aprendizaje supervisado	Algoritmos de aprendizaje supervisado es la rama de la informática que utiliza la experiencia pasada para aprender y utilizar su conocimiento para tomar decisiones futuras, tiene como objetivo generalizar un patrón detectable o crear una regla desconocida a partir de ejemplos dados(Dangeti, 2017).	Mediciones de rendimiento	- Matriz de confusión - Accuracy - Sensibilidad
Variable dependiente: Rendimiento académico	El rendimiento académico es un referente del nivel de aprendizaje logrado por el estudiante en el aula y es el primer objetivo de la educación. (Chúmbez Rodríguez, 2017).	Sociales	- Genero - Edad - Estado civil - Financiamiento de estudio
		Económicos	- Contexto socioeconómico - Trabajo
		Académicos	- Modalidad de ingreso - Promedio semestral - Tipo de preparación - Bienestar psicológico

1.6. Hipótesis

1.6.1. Hipótesis general

El rendimiento académico de los estudiantes del primer semestre de la carrera profesional Ingeniería de Sistemas e Informática de la UNAMAD se puede predecir mediante algoritmos de aprendizaje supervisado.

1.6.2. Hipótesis específicas

1) El rendimiento académico de los estudiantes del primer semestre de la carrera profesional de Ingeniería de Sistemas e Informática de la UNAMAD se puede predecir mediante indicadores sociales, económicos y académicos.
2) Los algoritmos de aprendizaje supervisado pueden predecir el rendimiento académico de los estudiantes del primer semestre de la carrera profesional de Ingeniería de Sistemas e Informática de la UNAMAD.

1.7. Justificación

El aprendizaje supervisado es una técnica utilizada para la extracción de conocimientos de un conjunto de datos empleando métodos y algoritmos. La Universidad Nacional Amazónica De Madre De Dios cuenta con grandes cantidades de datos de los estudiantes, docentes y administrativos. El procesamiento de estos datos puede generar nuevos conocimientos o información relevante para la toma de decisiones, la presente tesis emplea el uso de algoritmos matemáticos para la predicción del posible rendimiento del estudiante en sus actividades académicas mediante sus primeros datos de ingreso a la universidad, de igual manera el estudiante podrá conocer su posible rendimiento académico ayudándolo en su diagnóstico de sus posibles fortalezas, debilidades, amenazas u otros. Esta investigación dará un aporte importante al campo del conocimiento científico del aprendizaje automático con implicancia al rendimiento académico, de igual manera el estudio servirá de apoyo para la toma de decisiones de la universidad en las áreas académicas generando posibles intervenciones a los estudiantes en lo

económico, social, becas, incentivos, tutorías personalizadas, recomendaciones, apoyo psicológico u otros.

1.7.1. Justificación tecnológica

Con el uso de las nuevas tecnologías se pretende predecir el rendimiento académico de los estudiantes de la Carrera Profesional de Ingeniería de Sistema e Informática, es decir se busca encontrar el mejor algoritmo de aprendizaje supervisado para predecir el rendimiento académico y así obtener nuevos conocimientos de los datos almacenados en la Universidad Nacional Amazónica De Madre De Dios.

1.7.2. Justificación Cultural

Teniendo en cuenta la gran importancia que tiene la educación en nuestro país y las pocas investigaciones realizadas en este entorno. Es necesario el estudio debido a la gran relevancia social y pedagógica que tiene este campo de conocimientos.

1.7.3. Justificación Económica

La grande inversión por parte del estado peruano en el sector de educación y la baja calidad educativa son evidentes, afrontar los gatos y nuevos métodos de aprendizaje son indispensables, por otro lado, el desarrollo de este estudio pretende prevenir los gastos económicos por parte del estado peruano ayudando a tomar decisiones al momento de implementar nuevas técnicas para el óptimo aprendizaje de los estudiantes de la Universidad Nacional Amazónica De Madre De Dios.

1.7.4. Justificación Teórica

La gran cantidad de métodos de enseñanza impartida por los expertos o educandos tendrías mayor eficacia si son adecuados al curso y a los estudiantes, conociendo el futuro posible rendimiento de los estudiantes se podría realizar los ajustes necesarios en cada estudiante, mejorando así las técnicas de aprendizaje y buscando alcanzar los objetivos educacionales de la Universidad Nacional Amazónica De Madre De Dios, es decir se espera que este estudio provea de conocimientos que puedan ser utilizados durante la toma de decisiones de los directivos de la universidad.

1.8. Consideraciones éticas

Para el desarrollo del presente estudio se cumplirá con la privacidad de la información de los sujetos de estudio, es decir se mantendrá la confidencialidad durante la manipulación de información proporcionada por Universidad Nacional Amazónica de Madre de Dios, asimismo se hará uso de la norma ISO para garantizar las citas adecuadas por otro lado, la investigación estará siendo monitoreada por especialistas encargados de validar los diferentes procesos realizados en el estudio.

CAPITULO II. MARCO TEORICO

2.1. Antecedentes de estudio

2.1.1. Antecedentes internacionales

En la investigación de Tomasevic et al. (2020) se refiere que el gran aumento de datos disponibles de aprendizaje impulsó el desarrollo de la minería de datos educativos con el fin de comprender y optimizar el proceso de aprendizaje. El objetivo de la investigación fue proporcionar un análisis exhaustivo y una comparación de las técnicas de aprendizaje automático supervisado de última generación aplicadas para resolver la tarea de predicción del rendimiento de los exámenes de los estudiantes, es decir, descubrir a los estudiantes con un "alto riesgo" de abandonar el curso. y predecir sus logros futuros. Se aplicó la clasificación binaria para predecir si el estudiante aprobó o desaprobó el examen. La investigación concluyó que las funciones adecuadas de adquisición de datos y la interacción del estudiante con el entorno de aprendizaje es un requisito previo para garantizar una cantidad suficiente de datos para el análisis.

Feng (2019) realizó un estudio en la Universidad Central de Florida para predecir el rendimiento académico de los estudiantes con el árbol de decisiones y una red neuronal. El estudio aplicó varios métodos de clasificación para descubrir y encontrar patrones ocultos en la base de datos de los estudiantes, el objetivo fue demostrar cómo aplicar las técnicas estadísticas y de aprendizaje automático para predecir el éxito académico de los estudiantes. El conjunto de datos utilizado contiene las calificaciones, los sexos, las edades, los cursos históricos de los estudiantes, y el rendimiento final de los estudiantes se clasificó en cinco niveles: excelente, muy bueno, bueno, promedio y malo. Los resultados revelaron que los factores significativos que más influyen en el proceso de clasificación son el puntaje

de admisión a la universidad de los estudiantes y los exámenes universitarios de primer año.

Fernandes et al. (2019) en su investigación presentaron un análisis predictivo del desempeño académico de los estudiantes de las escuelas públicas del Distrito Federal de Brasil, para ello se realizó un análisis estadístico descriptivo de las notas de los estudiantes. El objetivo de la investigación fue obtener patrones implícitos a partir de los datos analizados empleando el método de clasificación de minería de datos y el algoritmo de aumento de gradiente. El resultado de la clasificación y la metodología CRIS-DM mostro que las variables notas, ausencia, asignatura escolar, vecindario y otros son de gran relevancia para realizar la predicción. La investigación concluye que los atributos identificados ayudaron durante la predicción, así también el estudio brindará información importante para orientar a coordinadores, docentes y gerentes en la toma de decisiones sobre el año escolar, cursos dictados entre otros.

Waheed et al. (2020) presentaron su investigación en la que se utilizaron redes neuronales artificiales profundas para procesar las características que se relacionan con el rendimiento académico de los estudiantes. La investigación aplico un análisis doble, para el primer análisis se minó las actividades de los estudiantes con el portal VLE y los datos demográficos estáticos, en el segundo análisis se extrajo datos de los clics trimestrales para cada estudiante en cada curso nuevo. El estudio tuvo como objetivo ayudar a los institutos a formular un marco necesario para el apoyo pedagógico, facilitando el proceso de toma de decisiones de educación superior hacia una educación sostenible. Los autores indicaron que los modelos propuestos obtuvieron el siguiente porcentaje de predicción la regresión logística 84% y la máquina de vectores 93%. La investigación concluye que es posible predecir el desempeño del estudiante mediante las redes neuronales artificiales profundas y los resultados obtenidos pueden ser usados para las pautas estudiantiles pedagógicas constructivas y formativas.

Xu et al. (2019) en su investigación desarrollada en la Facultad de Humanidades y Ciencias Sociales de la Universidad de Beihang, en China, plantearon tres objetivos i) investigar diferencias significativas entre los grupos

de rendimiento para los comportamientos de uso en línea, ii) identificar las características de uso de Internet que se correlacionan con el rendimiento académico de los estudiantes y iii) predecir el rendimiento del estudiante a partir de los datos de uso de Internet con modelos de aprendizaje supervisados. Esta investigación se centró en las calificaciones finales de los estudiantes en los cursos obligatorios registrados en el sistema de gestión de asuntos académicos de esa universidad, además se recopilaron datos de uso de internet del campus de la universidad de cada estudiante. Para realizar la predicción del rendimiento académico se usaron los árboles de decisiones, las redes neuronales y máquina de vectores de soporte, y para determinar si existe una relación entre el comportamiento de los estudiantes y el rendimiento académico se empleó el coeficiente de correlación de Spearman. Durante el desarrollo de la predicción del rendimiento académico se logró observar un mejor porcentaje de predicción por parte de la máquina de vectores de soporte. La investigación concluye que con los datos del comportamiento de uso de internet son datos suficientes para predecir el rendimiento académico con una alta precisión, y se determinó que el uso excesivo del internet en entornos no académicos aumenta las probabilidades de reprobar cursos de los semestres académicos, el uso académico del internet en entornos educativos ayuda a mejorar el desempeño académico de muchos estudiantes

Méndez Aguirre & Guillermo López (2019) en su investigación académica plantearon como objetivo evaluar la incidencia del uso de machine learning en la predicción del desempeño academico. Este estudio empleó Python como lenguaje de programación, Jupyter Lab y Pycharm como entorno de desarrollo, Sklearn para el trabajo con machine learning, librería PyMc3 para las redes bayesianas y algoritmos, Pandas para trabajar con el conjunto de datos, Numpy para hacer los cálculos matemáticos, Scipy para realizar los cálculos de alto nivel en matemática y Django para el desarrollo de la aplicación web. Por otro lado, hallo que el uso de la técnica de aprendizaje automático y el desarrollo de un modelo que pronostica el desempeño de los estudiantes permiten convertir datos sin procesar en inteligencia accionable. Se concluye que la intervención de los algoritmos matemáticos, el aprendizaje

automático, las simulaciones, la extracción, procesamiento y almacenamiento de datos provee un modelo estable.

2.1.2. Antecedentes nacionales

Menacho Chiok (2017) realizó un estudio en la Universidad Agraria La Molina, Lima – Perú, el cual tuvo como objetivo aplicar técnicas de minería de datos en el curso de estadística general y predecir la calificación final de los estudiantes. Utilizó herramientas tecnológicas como las redes neuronales, redes bayesianas, regresión logística, arboles de decisión y otros. La muestra estuvo constituida por 914 estudiantes matriculados en el curso de estadística general durante los semestres 2013-II hasta el semestre 2014-I. Se utilizaron técnicas de clasificación del aprendizaje supervisado y técnicas para evaluar los algoritmos de clasificación, una de estas herramientas es la matriz de confusión que nos permitirá conocer la distribución de la clasificación real y la clasificación predicha en las diferentes categorías, además se utilizó la técnica de área bajo la curva ROC que es un índice que mide el desempeño de la clasificación estableciendo ciertos parámetros de aceptación y por último se empleó la técnica de evaluación del coeficiente kappa. La investigación concluye que el algoritmo clasificador Naive de Bayes fue la mejor para predecir el rendimiento académico con una precisión de 71% durante la clasificación y además que se logró identificar la influencia de las variables en el curso de estadística general tales como la nota, promedio, situación del curso y otros.

La tesis de Holgado-Apaza (2018) tuvo como objetivo principal detectar los patrones de bajo rendimiento académico de los estudiantes de la Universidad Nacional Amazónica de Madre de Dios usando minería de datos, para ello se aplicaron las técnicas de clasificación, la metodología CRISP-DM y los algoritmos Random Forest, C5.O y CART. Se utilizó el lenguaje R para predecir e identificar las variables más influyentes en el modelo de clasificación para el bajo rendimiento académico de los estudiantes. La investigación concluye que la aplicación de los algoritmos CART y C5.O hicieron posible identificar el perfil de estudiantes con bajo rendimiento académico y se identificó las variables más influyentes con relación al

rendimiento académico tales como el servicio del comedor, la cantidad de cursos matriculados y la carrera profesional.

Yamao et al. (2018) plantearon como objetivo de la investigación predecir el rendimiento académico en el primer ciclo de los estudiantes ingresantes a la Escuela Profesional de Ingeniería en Computación y Sistemas de la Universidad de San Martín de Porres empleando minería de datos. La muestra estuvo constituida por los datos de tipo social, económico y académico de 1304 ingresantes en el periodo 2010 al 2015. Para realizar la predicción del rendimiento académico se usaron tres técnicas: regresión lineal, árbol de decisiones y máquina de soporte vectorial, el mejor resultado de 82.87% se obtuvo con el árbol de decisiones. Se concluye que con la minería de datos es posible predecir el rendimiento académico de los estudiantes, esto posibilita detectar a los estudiantes que podrían tener problemas en sus estudios durante el primer semestre; siendo los factores más influyentes la nota en el examen de admisión, género, edad, ingresos y distancia del hogar al centro de estudios.

La tesis de Orihuela Maita (2019) se planteó como objetivo predecir el rendimiento académico de los estudiantes de la Facultad de Ingeniería de Sistemas de la Universidad Nacional del Centro del Perú. El tipo de investigación es tecnológica, el nivel es explicativa causal y el diseño es pre experimental pues aplicando la ciencia de datos como estímulo se realizará las mediciones de los resultados obtenidos en la predicción del rendimiento académico de los estudiantes. La muestra estuvo conformada por los registros recopilados de los estudiantes de la carrera de Ingeniería de Sistemas desde el periodo académico 2016-I hasta el 2019-I. Se aplicaron técnicas de limpieza de datos, exploración y la aplicación de los modelos de aprendizaje supervisado de Machine Learning, Regresión Logística y el Random Forest. Se logró predecir el rendimiento académico de los estudiantes con un 80% de precisión para los datos de entrenamiento y un 76% para los datos de validación es decir el modelo de regresión logística obtuvo un 77% de precisión con los datos de entrenamiento y un 75% con datos de validación, asimismo Random forest obtuvo un 84% de precisión con los datos de entrenamiento y un 76% con los datos de validación. Se concluye que el uso

de la ciencia de datos se logró predecir el rendimiento académico de los estudiantes de la Facultad de Ingeniería de Sistemas de la Universidad Nacional del Centro del Perú.

2.2. Marco teórico

2.2.1. Rendimiento académico

El desempeño o rendimiento académico es un referente del nivel de aprendizaje logrado por el estudiante en el aula y es el primer objetivo de la educación. Existen diversas variables externas que intervienen en el desempeño académico, como la aptitud del docente, el entorno de clase, el entorno familiar, el plan educativo y otros. Es conveniente dejar en claro que el aprovechamiento académico no es equivalente al desempeño académico, ya que el responsable del desempeño académico es el estudiante, mientras que el aprovechamiento es la consecuencia del proceso de educación impartida por un docente o entidad. En la investigación se refiere que las variables que explican el rendimiento académico son: nivel socio económico, nivel socio cultural, aptitud del docente y aptitud de los padres de familia con el desempeño académico del estudiante (Chúmbez Rodríguez, 2017).

El desempeño académico de los estudiantes se ve perjudicado por muchos factores que incluyen: edad, género, inglés como segundo idioma, estado laboral, calificaciones de admisión, dentro del desempeño del programa, habilidades de pensamiento crítico, personalidad, autoeficacia y compromiso académico (Pitt et al., 2012)

Es improbable que los estudiantes con adicción a los teléfonos inteligentes logren un adecuado rendimiento académico, especialmente si usan aplicaciones de redes sociales durante el estudio o las actividades de asistencia a clases. Los administradores y el personal académico deben ser conscientes de cómo sus estudiantes usan la tecnología y el efecto perjudicial en su rendimiento académico, por lo que deben sensibilizar a los estudiantes sobre el uso racional del teléfono inteligente, especialmente durante aquellas actividades estrictamente relacionadas con su rendimiento académico. Los debates abiertos o los servicios de tutoría pueden ayudar a los estudiantes a

conocer estrategias para administrar su tiempo y su carga de trabajo académico. Finalmente, si se consideran las variables de género y departamento, los grupos con peor desempeño son las mujeres en el departamento de humanidades y los hombres en los campos científicos (Giunchiglia et al., 2018).

Existen múltiples factores asociados al rendimiento académico Garbanzo Vargas (2012), entre ellos:

- Determinantes personales. Son aquellos factores de carácter personal y comprende múltiples competencias tales como: formación académica previa a la universidad, nota de acceso a la universidad, competencia y condiciones cognitivas, aptitudes, bienestar psicológico, autoeficacia percibida, asistencia a clases, aptitudes, inteligencia, sexo, motivación, satisfacción y abandono con respecto a los estudios.
- Determinantes sociales. Son factores que ejercen una acción en la vida académica de los estudiantes, asimismo estos factores interactúan consigo mismo o con las instituciones académicas, entre ellos: variables demográficas, nivel educativo de la madre, contexto socioeconómico, nivel educativo de los progenitores o adultos responsables del estudiante, diferencias sociales y entorno familiar
- Determinantes institucionales. Son factores que influyen en el rendimiento y a su vez se interrelacionan con los factores personales, sociales o consigo misma, los factores son los siguientes: pruebas específicas de ingreso a la carrera, complejidad en los estudios, relación estudiante – profesor, condiciones institucionales, ambiente estudiantil, servicios institucionales de apoyo, elección de los estudios según interés del estudiante.

Otros autores establecen los siguientes factores asociados al rendimiento académico (Montero Rojas, Villalobos Palma y Valverde Bermúdez 2007).

- Factores institucionales. Son aquellos factores que influyen en el aprendizaje de los estudiantes y que tienen relación con los ambientes educativos de las instituciones, la carrera que estudia el estudiante, grupos de estudio, cantidad de libros en la biblioteca y horario de curso matriculados.

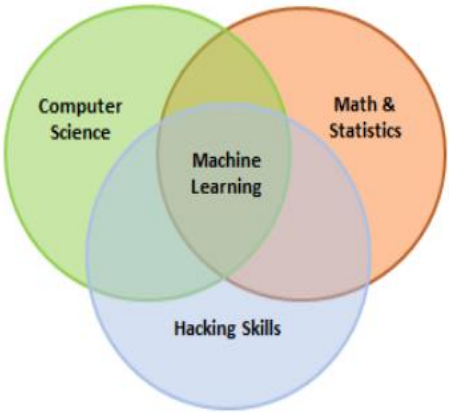
- Factores pedagógicos. Son aquellos factores que tienen gran relación con el docente que imparte catedra a los estudiantes, así mismo los métodos de enseñanza, programas, planes curriculares y el equipo docente son de gran importancia y cruciales para el rendimiento académico.
- Factores psicosociales. Son aquellos factores que consideran las relaciones de la sociedad con la persona, en los cuales se miden la peculiaridad personal de cada individuo mujer o varón, tales como autoestima, percepción del ambiente académico por parte estudiante, ansiedad, motivación y autoestima en entornos académicos.
- Factores sociodemográficos. Son aquellos factores que consideran el nivel educativo de los padres de familia, sexo, tipo de colegio en el que culmino su etapa escolar y el nivel socioeconómico familiar, así mismo es importante tener en cuenta el ambiente sociocultural que facilitan el acceso a movilidad, educación y empleo.

2.2.2. Aprendizaje Automático o Machine Learning (ML)

El aprendizaje automático es la rama de la informática que utiliza la experiencia pasada para aprender y utilizar su conocimiento para tomar decisiones futuras. Se encuentra en la intersección de la informática, la ingeniería y las estadísticas. Tiene como objetivo generalizar un patrón detectable o crear una regla desconocida a partir de ejemplos dados. (Dangeti, 2017).

Figura 1

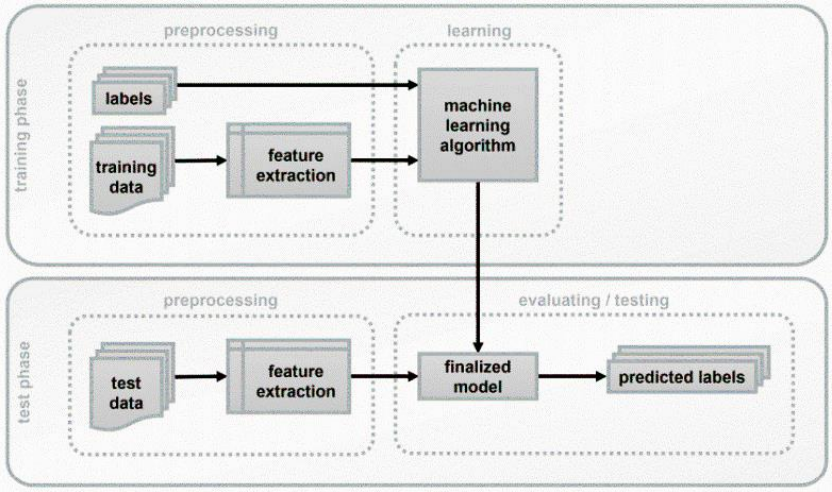
El aprendizaje automático



El aprendizaje automático se trata de construir modelos matemáticos para comprender los datos. El aspecto de aprendizaje ingresa a este proceso al otorgar a un modelo de aprendizaje automático la capacidad de ajustar sus parámetros internos para que el modelo explique mejor los datos. En cierto sentido, esto puede entenderse como el modelo que aprende de los datos. Una vez que el modelo ha aprendido lo suficiente, sea lo que sea que eso signifique, podemos pedirle que explique los datos recientemente observados (Beyeler, 2018).

Figura 2

Flujo de trabajo para plantear problemas de machine learning



El aprendizaje automático es un enfoque de ingeniería que otorga la máxima importancia a cada técnica que aumenta o mejora la propensión a cambiar de forma adaptativa. Su objetivo principal es estudiar, diseñar y mejorar modelos matemáticos que se puedan entrenar (una vez o de forma continua) con datos relacionados con el contexto (proporcionados por un entorno genérico), para inferir el futuro y tomar decisiones sin completar conocimiento de todos los elementos influyentes (factores externos). En otras palabras, un agente (que es una entidad de software que recibe información de un entorno, elige la mejor acción para alcanzar un objetivo específico y observa los resultados) adopta un enfoque de aprendizaje estadístico, tratando de determinar las distribuciones de probabilidad correctas y úselas para calcular la acción (valor o decisión) que es más probable que tenga éxito (con el menor error) (Bonaccorso, (2017).

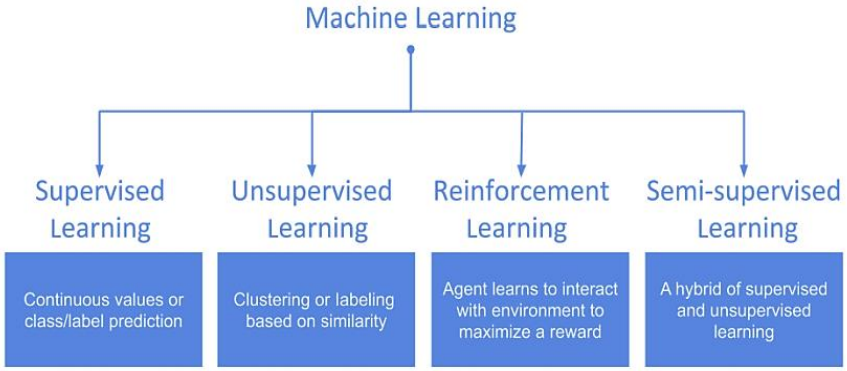
La tecnología de aprendizaje automático tiene un gran potencial para contribuir al desarrollo de algoritmos de visión flexibles y robustos, mejorando así el rendimiento de los sistemas de visión prácticos. Se espera que los sistemas de visión basados en el aprendizaje proporcionen un mayor nivel de competencia y una mayor generalidad. El aprendizaje puede permitirnos utilizar la experiencia adquirida en la creación de un sistema de visión para un dominio de aplicación a un sistema de visión para otro dominio mediante el desarrollo de sistemas que adquieran y mantengan el conocimiento. Afirmamos que el aprendizaje representa la próxima frontera desafiante para la investigación de la visión por computadora el aprendizaje automático ofrece métodos efectivos para la visión por computadora para automatizar los procesos de adquisición y actualización de modelos / conceptos, adaptando parámetros y representaciones de tareas, y utilizando la experiencia para generar, verificar y modificar hipótesis (Sebe et al., 2005).

Machine learning es perfecto para solucionar problemas que necesitan múltiples ajustes o gran cantidad de reglas (es ideal para sintetizar el código y tiene un mejor funcionamiento). Pueden solucionar problemas difíciles que un enfoque tradicional no puede: el aprendizaje automático puede dar una posible solución a problemas complejos. Un sistema de aprendizaje automático es capaz adecuarse a datos recientes. Puede conseguir información de problemas complicados y también de grandes cantidades de datos (Gron, 2019).

Los algoritmos de ML se dividen en tres categorías principales y otra forma que se deriva de ellos, en función de los datos de entrada que reciben y el tipo de salida que se supone que producen. Siendo los siguientes: aprendizaje supervisado, aprendizaje no supervisado, aprendizaje por reforzamiento y aprendizaje semi-supervisado (Gron, 2019).

Figura 3

Tipos de aprendizaje automático



a) Aprendizaje supervisado

En esta forma de ML el algoritmo se presenta con una gran cantidad de muestras de entrenamiento, que contiene información sobre todos los parámetros o características que se utilizarían para determinar una característica de salida. Esta característica de salida podría ser un rango continuo de valores o una colección discreta de etiquetas. En base a esto, los algoritmos de ML supervisados se dividen en dos partes:

- Clasificación. Algoritmos que producen etiquetas discretas en la función de salida, como normal y no normal o un conjunto de categorías.
- Regresión. Cuando la función de salida tiene valores reales, por ejemplo, el número de votos que un partido político podría recibir en una elección, o la temperatura de un material en el que se prevé que alcance su punto de fusión.

Algunos de los algoritmos de aprendizaje supervisado más conocidos son la regresión lineal, la regresión logística, las máquinas de vectores de soporte y los vecinos k más cercanos (Singh & Paul, 2020).

b) Aprendizaje no supervisado

El aprendizaje no supervisado se trata de modelar datos que vienen sin las etiquetas o respuestas correspondientes. Con suficientes datos, puede ser posible encontrar patrones y estructura. Dos de las herramientas más poderosas que utilizan los profesionales del aprendizaje automático para aprender solo de los datos son la agrupación y la reducción de la dimensionalidad (Shukla, 2017).

Algunos de los algoritmos más importantes de aprendizaje no supervisado (Géron, 2017).

- Agrupación: K-medias, DBSCAN, análisis jerárquico de conglomerados (HCA).
- Detección de anomalías y novedad: SVM de una clase, bosque de aislamiento.
- Visualización y reducción de dimensionalidad: Kernel PCA, incrustación localmente lineal (LLE), análisis de componentes principales (PCA), incrustación de vecinos estocásticos distribuidos en t (t-SNE).
- Aprendizaje de reglas de asociación: a priori, ECLAT.

c) Aprendizaje por refuerzo

Estas técnicas han ganado mucha atracción en los últimos años en la aplicación de inteligencia artificial. En el aprendizaje por refuerzo, se deben tomar decisiones secuenciales en lugar de la toma de decisiones de una sola vez, lo que hace que sea difícil capacitar a los modelos en algunos casos Dangeti (2017).

Tiene como objetivo elegir la acción que produzca la mayor recompensa, dado un conjunto de datos de entrada que describe un contexto o entorno. Es dinámico e interactivo: el flujo de recompensas positivas y negativas impacta el aprendizaje del algoritmo, y las acciones tomadas ahora pueden influir tanto en el medio ambiente como en las recompensas futuras. La compensación entre la explotación de un curso de acción que se ha aprendido a dar una cierta recompensa y la exploración de nuevas acciones que pueden aumentar la recompensa en el futuro da lugar a un enfoque de prueba y error. El aprendizaje por refuerzo optimiza el aprendizaje del agente mediante la teoría de sistemas dinámicos y, en particular, el control óptimo de los procesos de decisión de Markov con información incompleta (Jansen, 2018).

d) Aprendizaje semi-supervisado

El aprendizaje semi-supervisado se ha utilizado con éxito para producir resultados más eficientes cuando se agregan algunas muestras etiquetadas a un problema que pertenece por completo al aprendizaje no supervisado. Además, dado que solo se etiquetan unas pocas muestras, se evita la complejidad del aprendizaje supervisado. Con este enfoque, podemos

producir mejores resultados de los que obtendríamos de un sistema de aprendizaje puramente no supervisado e incurrir en un menor costo computacional que un sistema de aprendizaje supervisado puro (Singh & Paul, 2020).

En esta investigación se emplearán los algoritmos supervisados dado que se cuenta con los sets de datos de los estudiantes, se tiene los datos de ingreso (variable independiente) de los estudiantes y sus determinadas calificaciones buenas o malas de cada estudiante (variable dependiente).

Para mostrar los datos de nuestra predicción se usará los algoritmos de clasificación ya que nuestras predicciones serán valores categóricos.

2.2.3. Algoritmos de aprendizaje supervisado

Algunos de los algoritmos de aprendizaje supervisado más importantes: Redes neuronales, Regresión logística, Máquinas de vectores de soporte (SVM), Bosques al azar y Árboles de decisión, Regresión lineal, Vecinos k-más cercanos (Gron, 2019).

a) Vecinos k-más cercanos (KNN)

Este algoritmo pertenece a una familia particular llamada instancia basada (la metodología se llama aprendizaje basado en instancia). Se diferencia de otros enfoques porque no funciona con un modelo matemático real. Por el contrario, la inferencia se realiza mediante la comparación directa de nuevas muestras con las existentes (que se definen como instancias). KNN es un enfoque que puede emplearse fácilmente para resolver problemas de agrupamiento, clasificación y regresión (Bonaccorso, 2018).

Los métodos de vecino más cercano se basan en una idea simple, se considera el conjunto de entrenamiento como modelo y hacemos predicciones sobre nuevos puntos en función de lo cerca que estén de los puntos del conjunto de entrenamiento. Pero como la mayoría de los conjuntos de datos contienen un grado de ruido, un método más común sería tomar un promedio ponderado de un conjunto de k vecinos más cercanos (Mcclure, 2017).

b) Regresión lineal

La regresión lineal puede ser uno de los algoritmos más importantes en estadística, aprendizaje automático y ciencias en general. Es uno de los

algoritmos más utilizados y es muy importante comprender cómo implementarlo y sus diversos usos. Una de las ventajas que tiene la regresión lineal sobre muchos otros algoritmos es que es muy interpretable. Terminamos con un número para cada característica que representa directamente cómo esa característica influye en el objetivo o la variable dependiente (Mcclure, 2017).

El objetivo de la regresión lineal es modelar la relación entre una o varias características y una variable objetivo-continua, dando a las computadoras la capacidad de aprender de los datos, el análisis de regresión es una subcategoría del aprendizaje automático supervisado. A diferencia de la clasificación, otra subcategoría del aprendizaje supervisado, el análisis de regresión tiene como objetivo predecir los resultados en una escala continua en lugar de las etiquetas de clase categóricas Raschka & Mirjalili, (2017).

c) Regresión logística

La regresión logística es una forma de convertir la regresión lineal en una clasificación binaria. Esto se logra transformando la salida lineal en una función sigmoidea que escala la salida entre cero y 1. El objetivo es un cero o 1, lo que indica si un punto de datos está o no en una clase u otra. Como estamos prediciendo un número entre cero o 1, la predicción se clasifica en el valor de clase “1" si la predicción está por encima de un valor límite especificado y la clase “0” en caso contrario (Mcclure, 2017).

d) Máquinas de vectores de soporte (SVM)

Los clasificadores SVM son modelos binarios o discriminantes, que trabajan en dos clases de diferenciación. Su principal tarea consiste básicamente en discriminar nuevas observaciones entre dos clases. Durante la fase de aprendizaje, estos clasificadores proyectan las observaciones en un espacio multidimensional llamado espacio de decisión y construyen una superficie de separación llamada límite de decisión que divide este espacio en dos áreas de pertenencia. En el caso más simple, es decir, el caso lineal, el límite de decisión estará representado por un plano (en 3D) o por una línea recta (en 2D). En casos más complejos, las superficies de separación son formas curvas con formas cada vez más articuladas. La SVM se puede utilizar tanto

en regresión, con Regresión Vectorial de Soporte, como en clasificación, con Clasificación de Vectores de Soporte (Nelli, 2018).

e) Árboles de decisión y bosques al azar

Son algoritmos de aprendizaje automático mudables los cuales realizan diferentes tareas como regresión o clasificación, estos algoritmos son eficaces con set de datos muy difíciles de procesar.

Los árboles de decisión también son los componentes fundamentales de los bosques aleatorios, que se encuentran entre los algoritmos de aprendizaje automático más potentes disponibles en la actualidad (Géron, 2017).

Son modelos semejantes a un proceso de decisión jerárquico estándar. En la mayoría de los casos, se emplea una familia especial, llamada árboles de decisión binarios, ya que cada decisión produce solo dos resultados. Este tipo de árbol es a menudo la opción más simple y razonable y el proceso de capacitación (que consiste en construir el árbol en sí) es muy intuitivo. La raíz contiene todo el conjunto de datos (Bonaccorso, 2018).

f) Redes neuronales

Una red neuronal es básicamente una secuencia de operaciones aplicadas a una matriz de datos de entrada. Estas operaciones suelen ser colecciones de sumas y multiplicaciones seguidas de aplicaciones de funciones no lineales (Mcclure, 2017).

Las redes neuronales son una clase de modelos matemáticos que se entrenan para producir y optimizar una definición para una función (o distribución) sobre un conjunto de características de entrada. El operador puede definir el objetivo específico de una aplicación de red neuronal dada utilizando una medida de rendimiento (típicamente una función de costo); de esta manera, las redes neuronales pueden usarse para clasificar, predecir o transformar sus entradas (Hearty, 2016).

2.2.4. Minería de datos

El data mining es procesar datos para aislar patrones y establecer relaciones entre entidades de datos dentro del conjunto de datos (Hearty, 2016).

2.2.5. Metodología CRISP-DM (Cross-Industry Standard Process for Data Mining)

CRISP-DM proporciona un ciclo de vida para proyectos de minería de datos, de igual manera establece las fases, tareas y sus relaciones entre las diferentes tareas. El ciclo de vida de un proyecto de minería de datos se compone de seis fases (Pete et al., 2000).

a) **Comprensión empresarial.** En esta fase se establece los requerimientos del proyecto y objetivos desde el punto de vista comercial, el planteamiento de esta fase ayudara a establecer un plan preliminar para lograr cada objetivo trazado.

b) **Comprensión de datos.** Para un adecuado procesamiento se inicia con tarea de recopilación de los datos y se sigue con las diferentes actividades para conocer los tipos de datos, calidad de datos y sus características, así mismo podremos establecer las hipótesis necesarias para conocer la información oculta en los datos.

c) **Preparación de datos.** Esta fase establece las actividades necesarias para determinar los conjuntos de datos finales para ser utilizados en los modelos de predicción, clasificación y otros. La fase de preparación puede realizarse muchas veces hasta encontrar los atributos, tablas y registros primordiales, de igual manera se realiza la limpieza de datos y su transformación para aplicar los diferentes modelos de aprendizaje automático.

d) **Modelado.** En esta fase se aplican las técnicas de modelado de los diferentes tipos de aprendizaje automático o supervisado, hasta obtener el menor incide de error posible.

e) **Evaluación.** La fase de evaluación aplica ciertas tareas para evaluar el índice de error del modelo propuesto, así mismo esta fase se encarga de lograr los objetivos comerciales propuestos y determinar los objetivos no logrados, por otro lado, la fase de evaluación debe determinar las decisiones sobre el uso de los resultados de la minería de datos.

f) **Despliegue.** Esta fase presenta al cliente los conocimientos adquiridos de los datos y establece los mecanismos adecuados para su uso del modelo entrenado, el propósito del modelo es agregar nuevos conjuntos de datos para aumentar la robustez del modelo.

2.3. Definición de términos

2.3.1. Lenguaje de programación Python

Se define comúnmente como un lenguaje de scripting orientado a objetos, una definición que combina el soporte para POO con una orientación general hacia los roles de scripting (Mark 2009).

2.3.2. Librería Numpy

Es una de las bibliotecas más importantes en Python para los cálculos numéricos. Añade soporte al núcleo de Python para matrices multidimensionales (y matrices) y operaciones vectorizadas rápidas en estas matrices (Sarkar et al., 2018).

2.3.3. Librería Pandas

Es una biblioteca de Python importante para la manipulación, discusión y análisis de datos, asimismo permite trabajar con datos transversales y datos basados en series de tiempo (Sarkar et al., 2018).

2.3.4. Librería sScikit-learn

Es una librería importante para la ciencia de datos y el aprendizaje automático ya que implementa una amplia gama de algoritmos de aprendizaje automático que cubren las principales áreas del aprendizaje automático, como clasificación, agrupación en clustering, regresión, etc. (Sarkar et al., 2018)**.**

2.3.5. Librería Matplotlib

Esta librería proporciona interfaces y herramientas para producir visualizaciones de calidad (Sarkar et al., 2018).

2.3.6. Librería Seaborn

Proporciona cuadrículas de facetas que nos ayudan a visualizar un mayor número de variables en gráficos bidimensionales (Sarkar et al., 2018).

2.3.7. Librería TensorFlow

Es una biblioteca más compleja para el cálculo numérico distribuido. Hace posible entrenar y ejecutar redes neuronales muy grandes de manera eficiente al distribuir los cálculos en potencialmente cientos de servidores multi-GPU (unidad de procesamiento de gráficos). TensorFlow se creó en Google y es compatible con muchas de sus aplicaciones de aprendizaje automático a gran escala. Fue de código abierto en noviembre de 2015 y la versión 2.0 se lanzó en septiembre de 2019 (Géron, 2017).

2.3.8. Librería Keras

Es una API de aprendizaje profundo de alto nivel que hace que sea muy sencillo entrenar y ejecutar redes neuronales. Puede ejecutarse sobre TensorFlow, Theano o Microsoft Cognitive Toolkit (anteriormente conocido como CNTK). TensorFlow viene con su propia implementación, llamada tf.keras, que brinda soporte para algunas funciones avanzadas de TensorFlow (Géron, 2017).

2.3.9. Anaconda

Anaconda hace que sea muy fácil lidiar con las diversas versiones de paquetes y actualizar los paquetes de dependencia o los paquetes dependientes, por otro lado, anaconda es una distribución libre para el lenguaje Python o R (Singh & Paul, 2020).

2.3.10. IDE Spyder

Es un entorno de desarrollo integrado (IDE) gratuito que viene con anaconda, que incluye edición, prueba y depuración en una única GUI (Naik & Oza, 2019).

CAPÍTULO III. METODOLOGÍA DE LA INVESTIGACIÓN

3.1. Tipo de estudio

La presente investigación es aplicada con enfoque cuantitativo ya que es un tipo de investigación en la que se utilizan técnicas y métodos estadísticos para analizar datos numéricos y obtener resultados cuantitativos. Esta investigación se centra en la recopilación y análisis de datos numéricos y suele ser utilizada para evaluar la efectividad de una intervención o para responder a una pregunta específica de investigación. El enfoque es cuantitativo ya que esta investigación busca evaluar la relación causal o determinar la existencia de una relación entre dos variables a través de técnicas y métodos estadísticos para analizar datos numéricos y obtener resultados cuantitativos (Arias, 2012).

3.2. Diseño del estudio

La presente investigación plantea un diseño no experimental de tipo correlacional puesto que se buscará conocer el grado de relación entre dos o más categorías o variables, en este caso las dimensiones socioeconómica, institucional o social y el rendimiento académico (Roberto Hernandez Sampie, 2014).

3.3. Población y muestra

3.3.1. Población

La población está conformada por las unidades o sujetos que serán analizados, contiene todos los elementos que cumplen con las propiedades establecidas por el investigador para ser estudiados (del Cid et al., 2011). En esta investigación la población de estudio estuvo conformada por los 861 registros de los estudiantes ingresantes a la carrera profesional de Ingeniería

de Sistemas e Informática de la UNAMAD durante los semestres académicos 2010-1 al 2020-2.

3.3.2. Muestra

La muestra es una fracción representativa e ilimitada que se separa de la población por el investigador para un estudio, con un margen de error y asimismo se pluraliza los resultados obtenidos a la población restante no seleccionada (Arias, 2012). En este estudio la muestra utilizada fue censal, es decir, que se usaron todas las unidades de investigación constituida por los 861 registros de los estudiantes ingresantes a la carrera profesional de Ingeniería de Sistemas e Informática de la UNAMAD durante los semestres académicos 2010-1 al 2020-2.

3.4. Métodos y técnicas

3.4.1. Métodos

La base de datos de los estudiantes de los ingresantes a la carrera profesional de Ingeniería de Sistemas e Informática en los semestres académicos 2010-I hasta 2020-II, fue proporcionada por la Dirección Universitaria de Asuntos Académicos (DAA) de la Universidad Nacional Amazónica de Madre de Dios.

3.4.2. Técnicas

Los datos para el análisis de las variables de estudio fueron obtenidos de la base datos a cargo de la Dirección de Asuntos Académicos, la cual fue solicitada formalmente para efectos de esta investigación. Dado que los datos pertenecen a una fuente secundaria no fue necesario un instrumento para recabar la información.

3.5. Tratamiento de los datos

En este apartado se formula las múltiples operaciones a las que serán sujetas los datos Arias, (2012). Para el tratamiento de los datos se aplicará el Chi cuadrado de Pearson, coeficiente de contingencia de Cramer y el coeficiente de contingencia de Pearson para conocer el nivel de relación entre las diferentes dimensiones asociados al rendimiento académico. De igual manera se utilizará la metodología CRISP-DM, el lenguaje de programación Python y

los algoritmos de aprendizaje supervisado: árbol de decisiones, K-NN y Naive Bayes.

CAPITULO IV: RESULTADO Y DISCUSIÓN

El presente estudio se desarrolla aplicando la metodología CRISP-DM, usada para procesos de minería de datos. Se basa en seis fases: comprensión del negocio, comprensión de datos, preparación de datos, modelado, evaluación y despliegue.

1) Comprensión del negocio

El rendimiento académico es un aspecto clave en la actualidad para mejorar la toma de decisiones en diferentes áreas administrativas. En este sentido, la predicción del rendimiento académico a través de los algoritmos de aprendizaje supervisado puede ser un factor crucial para mejorar el rendimiento académico.

2) Comprensión de datos

El conjunto de datos utilizado contiene información recopilada por la Dirección de Asuntos Académicos de la Universidad Nacional Amazónica de Madre de Dios. Este data set se publicó en el sitio web de Kaggle, al cual se puede acceder mediante el link https://www.kaggle.com/dsv/4462348.

Tabla 2

Descripción de atributos de datos

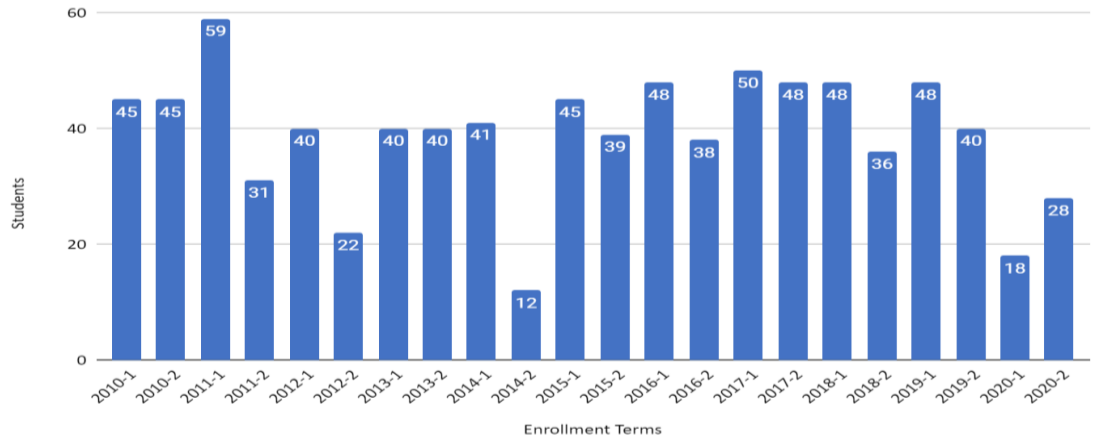
N°	Atributo	Descripción	Tipo
1	Tipo de admisión	Tipo de admisión del estudiante	Polinominal
2	Dependencia del estudiante	Dependencia estudiantil	Polinominal
4	Sexo	Género del estudiante	Binomial
5	Años	Edad de los estudiantes	Numérico
6	Estado civil	Estado civil del estudiante	Binomial
7	Preparación universitaria	Tipo de preparación estudiantil del estudiante	Polinominal
8	Bienestar psicológico	Bienestar psicológico del estudiante	Binomial
9	Condición de trabajo del estudiante	Situación laboral del estudiante	Polinominal
10	Socioeconómico	Situación Socioeconómica del estudiante	Polinominal
11	Calificación promedio ponderado	Promedio semestral del estudiante	Numérico

Ingresantes a la carrera profesional de Ingeniería de Sistemas e Informática distribuido por cada semestre académico

La cantidad total asciende a un total de 861 de estudiantes ingresantes desde el semestre académico 2010-I hasta el 2020-II.

Figura 4

Cantidad de Ingresantes

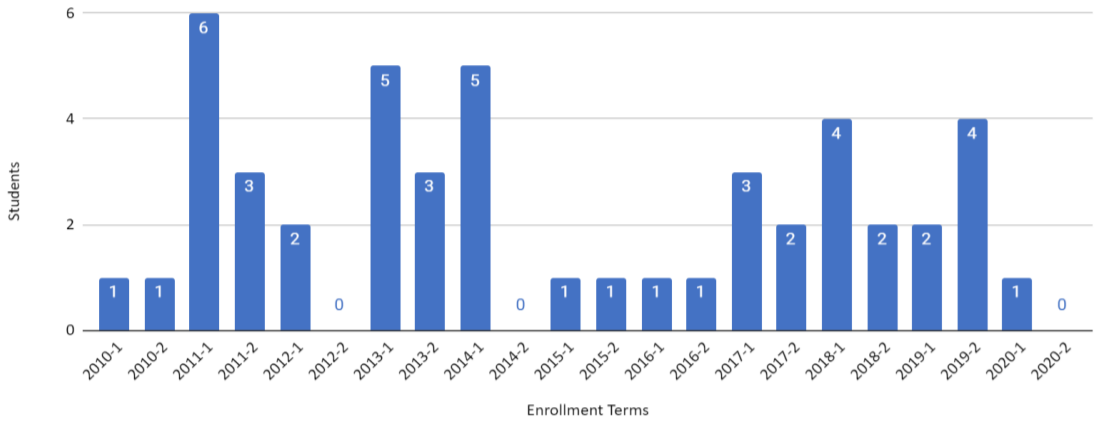


Ingresantes de la carrera profesional de Ingeniería de Sistemas e Informática que no cuentan con información

La cantidad de estudiantes que no cuentan con información totalizan 48 registros.

Figura 5

Cantidad de Ingresantes sin información

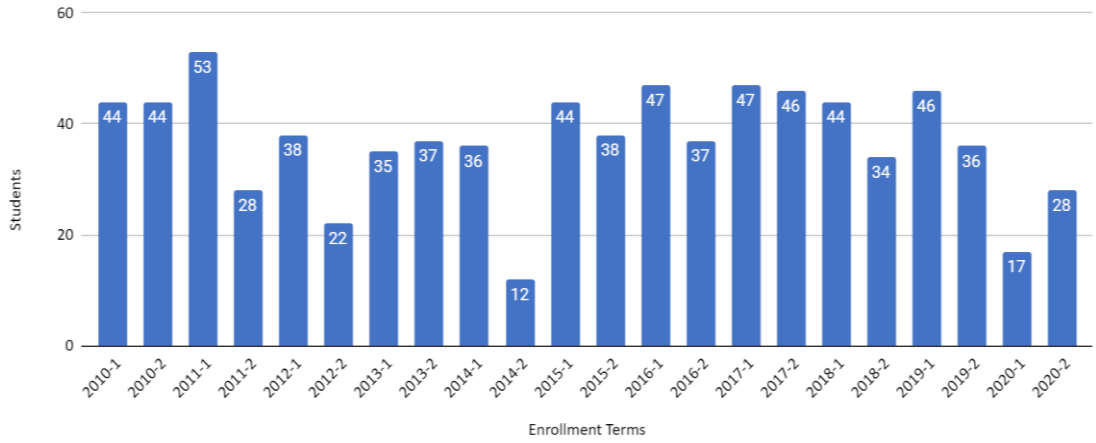


Ingresantes a la carrera profesional de Sistemas e Informática que cuentan con información

La cantidad de estudiantes que cuentan con información suman 813 registros.

Figura 6

Cantidad de Ingresantes con registros



Tipo de admisión del estudiante

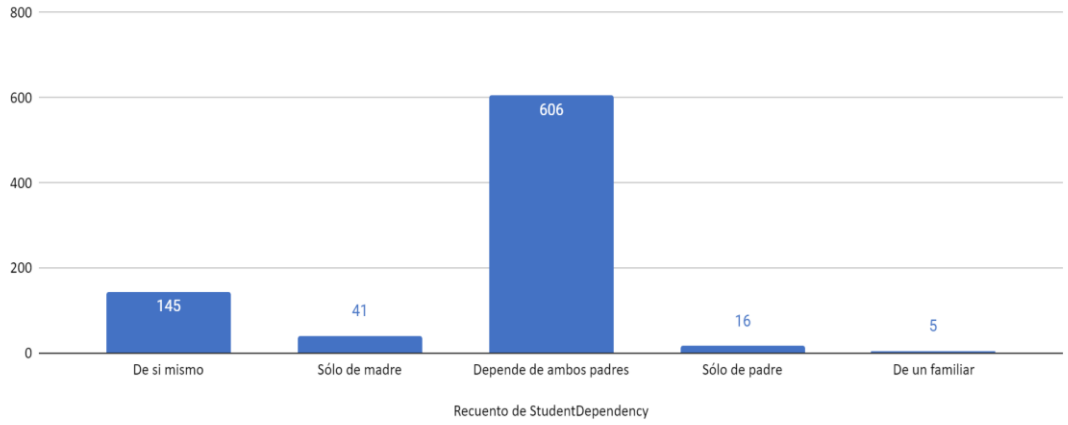
Figura 7

Cantidad de ingresantes por tipo de admisión

Dependencia del estudiante

Figura 8

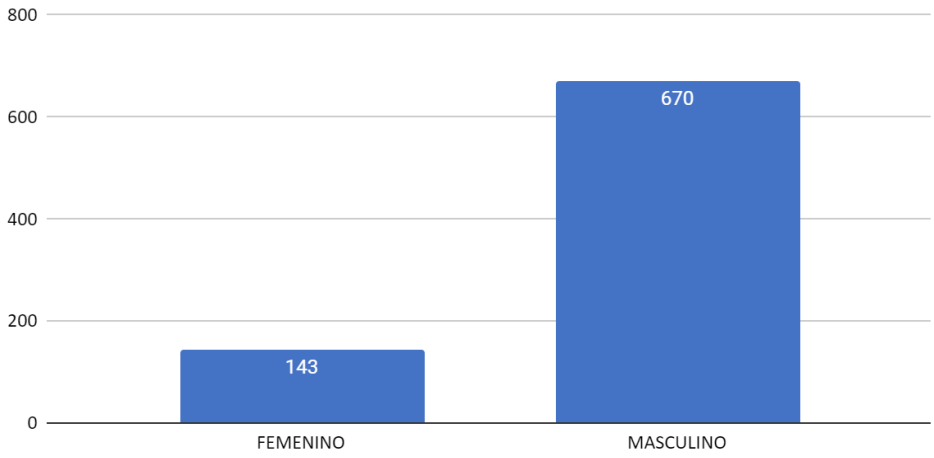
Cantidad de estudiantes según dependencia



Género del estudiante

Figura 9

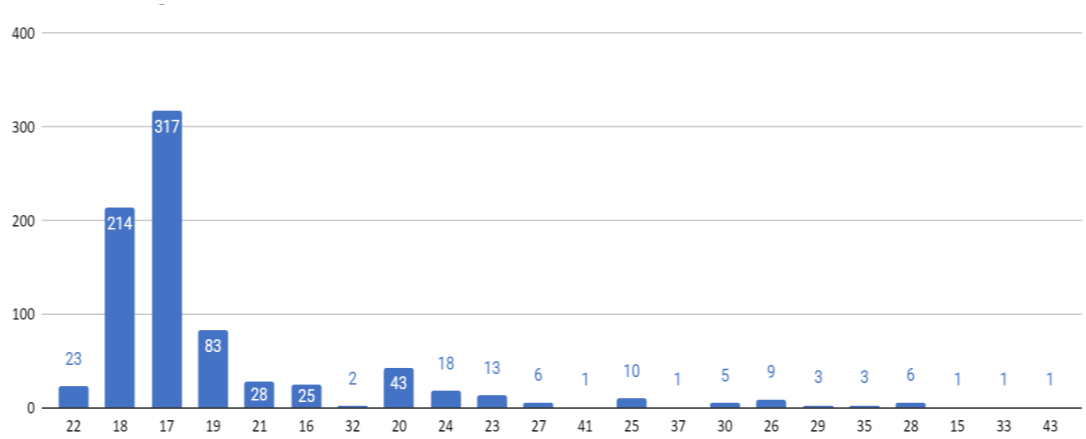
Cantidad de estudiantes según su género



Edad del estudiante

Figura 10

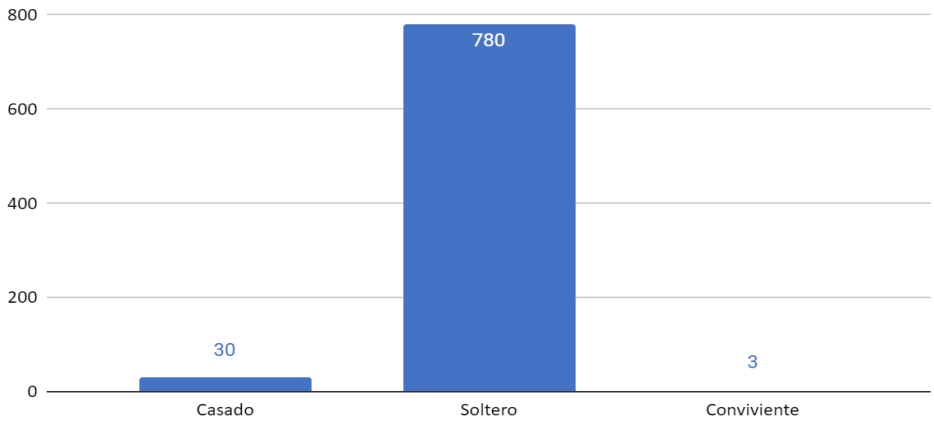
Cantidad de estudiantes según su edad



Estado civil del estudiante

Figura 11

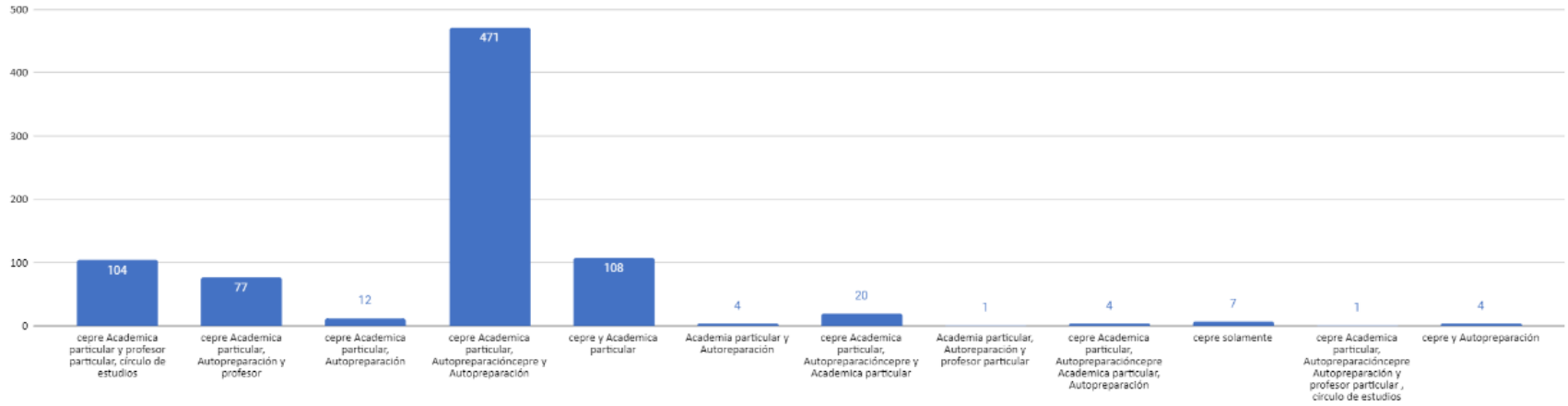
Cantidad de estudiantes de acuerdo a su estado civil



Tipo de preparación estudiantil del ingresante

Figura 12

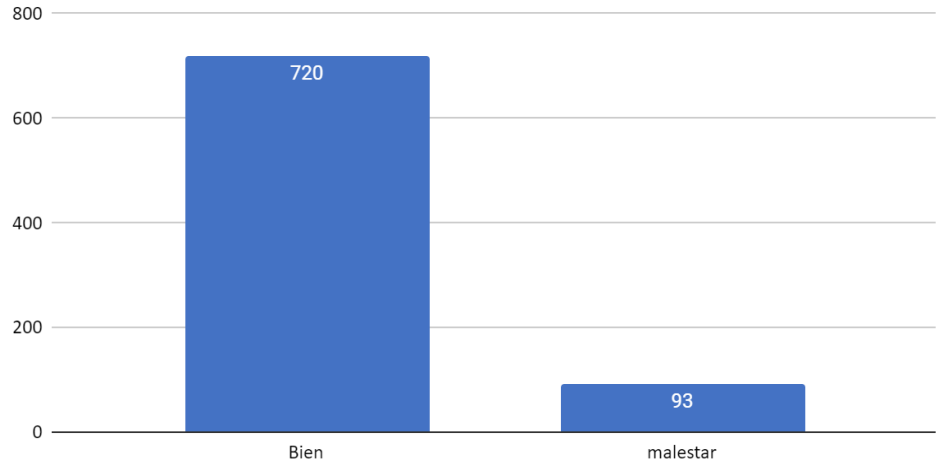
Cantidad de estudiantes de acuerdo con la preparación del estudiante



Bienestar psicológico del estudiante

Figura 13

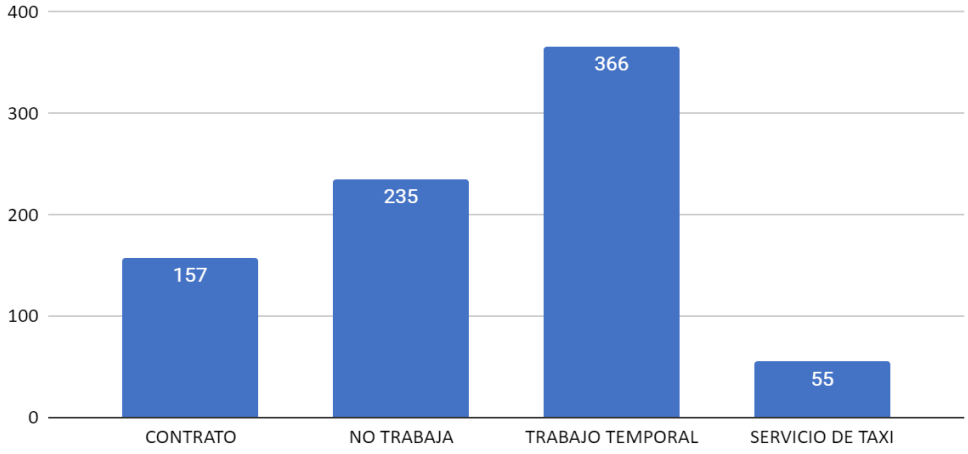
Cantidad de estudiantes según bienestar psicológico del estudiante



Situación laboral del estudiante

Figura 14

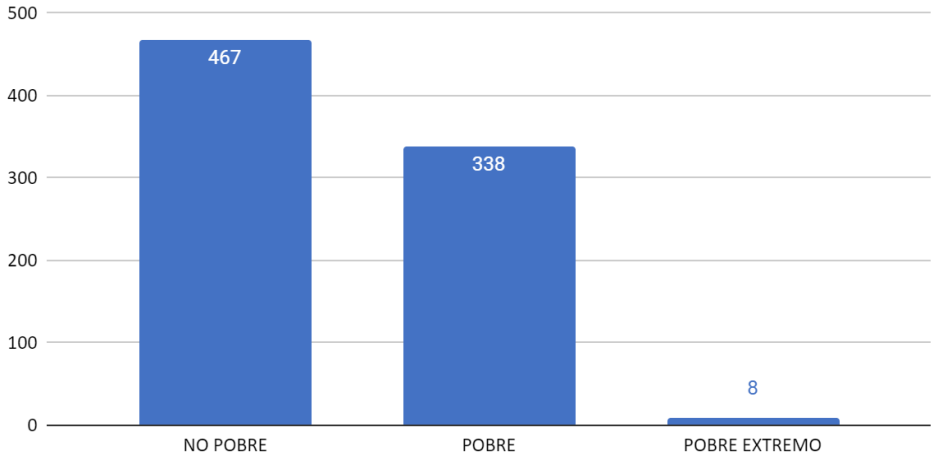
Cantidad de estudiantes según su situación laboral



Situación socioeconómica del estudiante

Figura 15

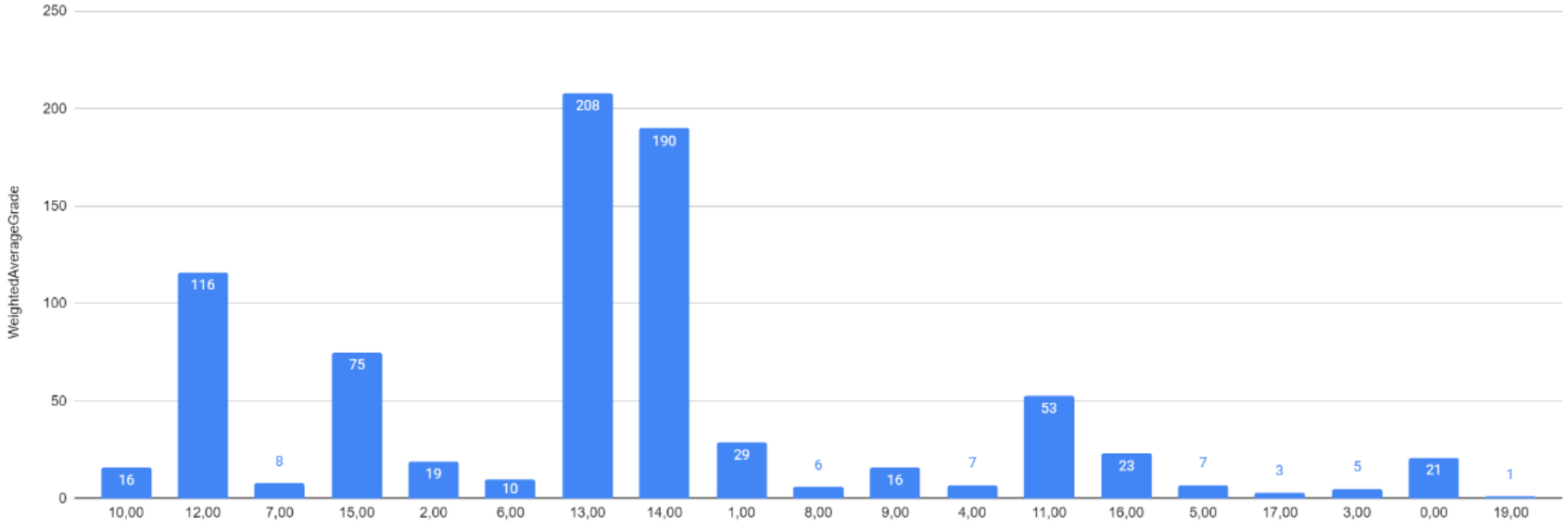
Cantidad de estudiantes según su situación socioeconómica



Promedio semestral del estudiante

Figura 16

Cantidad de estudiantes de acuerdo a su promedio semestral



3) Preparación de datos

En esta fase de la metodología, el conjunto de datos pasa por procesos de limpieza y transformación de datos para estar listo para las próximas etapas del ciclo de vida de CRISP-DM, En este sentido, el primer paso fue buscar instancias duplicadas, valores faltantes y valores atípicos, Se encontró instancias duplicadas, por lo cual se procedió a hacer el drop de los siguientes atributos como situación socioeconómica, dependencia estudiantil, situación civil, bienestar psicológico y situación laboral, solo fue admitido los registros del semestre de ingreso de cada estudiante, por otro lado, no se encontró valores faltantes o valores atípicos.

luego, a través de la evaluación de los atributos se pudo observar los pesos respecto al Promedio ponderado semestral del estudiante, asimismo con este análisis se detectó los atributos con mayor relación (peso) los cuales son la dependencia estudiantil y la Situación Socioeconómica del estudiante respecto al Promedio semestral del estudiante.

Tabla 3

Correlación de atributos de los datos en estudio

N	Atributos	Calificación promedio ponderada
1	Dependencia del estudiante	153.482
2	Socioeconómico	136.245
3	Sexo	58.816
4	Bien psicológico	-1.672
5	Estado civil	-22.913
6	Condición de trabajo del estudiante	-92.197
7	Edad	-100.916
8	Tipo de admisión	-137.798
9	Preparación Universitaria	-154.236

Resumen estadístico

Tabla 4

Resumen estadístico de atributos

	Tipo de admisión	Dependencia del estudiante	Sexo	Años	Estado civil	Preparación Universitaria	Bien psicológico	Condición de trabajo del estudiante	Socioeconómico	Calificación promedio ponderado
count	813.000.000	813.000.000	813.000.000	813.000.000	813.000.000	813.000.000	813.000.000	813.000.000	813.000.000	813.000.000
mean	3.611.316	821.648	1.175.892	18.809.348	47.970	9.889.299	114.391	1.974.170	1.435.424	11.679.410
std	1.619.302	1.538.160	380.962	3.134.004	260.563	6.351.374	318.482	1.005.194	515.594	3.920.685
min	1.000.000	0	1.000.000	15.000.000	0	0	0	1.000.000	1.000.000	0
25%	2.000.000	0	1.000.000	17.000.000	0	3.000.000	0	1.000.000	1.000.000	11.520.000
50%	5.000.000	0	1.000.000	18.000.000	0	15.000.000	0	2.000.000	1.000.000	12.940.000
75%	5.000.000	1.000.000	1.000.000	19.000.000	0	15.000.000	0	3.000.000	2.000.000	13.860.000
max	8.000.000	4.000.000	2.000.000	43.000.000	3.000.000	15.000.000	1.000.000	4.000.000	3.000.000	19.000.000

4) Evaluación

En esta etapa se pretende evaluar, para cada escenario, el desempeño de los diferentes clasificadores. La evaluación se basó en la matriz de confusión, que es una tabla de clasificación. Por otro lado, también se evalúo las mediciones de rendimiento: matriz de confusión, Accuracy y precisión.

La precisión, se refiere a la proporción entre las instancias correctamente clasificadas y todas las instancias clasificadas respecto al pro medio semestral. Comprueba si el DMM (Data Mining Models) puede detectar a los estudiantes con un buen rendimiento y se puede definir como la relación entre el rendimiento académico y las instancias clasificadas como pertenecientes a la clase positiva.

4.1. Resultados y discusión

El modelamiento de clasificación predictiva consiste en aproximar una función de mapeo de las variables de entrada a variables de salida discretas. Previo al análisis de datos mediante modelos de machine learning se realizó la preparación de datos, los cuales servirán como entrada al modelo. El conjunto de datos se dividió en un 75% para entrenamiento y un 25% para prueba con la finalidad de entrenar el modelo y evaluar el rendimiento. Luego, se emplearon tres modelos de aprendizaje supervisado: K-NN, Naive Bayes y árbol de decisiones para predecir la clasificación del rendimiento académico usando las dimensiones identificadas. Para las simulaciones se utilizó Google Colab con el lenguaje de programación Python.

Para el modelo K-Vécinos más cercanos las métricas de rendimiento son mostradas en la **tabla 5**, asimismo de la evaluación realizada se observó que la precisión del modelo supera el 80%, por otro lado, el modelo no logra encontrar todas las instancias de una clase especifica, asimismo se observa que la calidad de predicción del modelo es un poco baja y la calidad del modelo supera el 50%.

Figura 17

Matriz de confusión del modelo K-Vécinos más cercanos

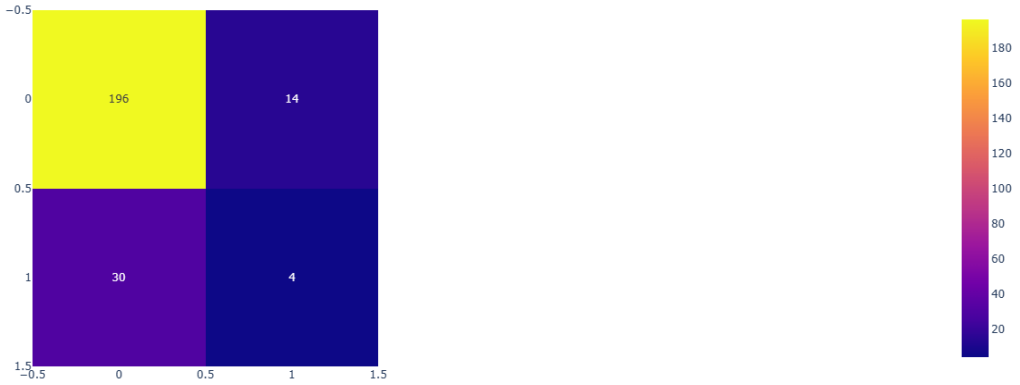


Tabla 5

Métricas de evaluación del clasificador K-Vecinos más cercanos

N°	Métricas de evaluación del Algoritmo	KNN
1	accuracy	0.819672
2	Recuperación	0.117647
3	Precisión	0.222222
4	roc_auc_score	0.525490

El modelo de Naïve Bayes no tuvo una buena precisión, por otro lado, el modelo si logra encontrar las instancias de una clase especifica, la calidad de precisión es muy baja y la calidad del modelo supera el 50%.

Figura 18

Matriz de confusión del modelo Naïve Bayes

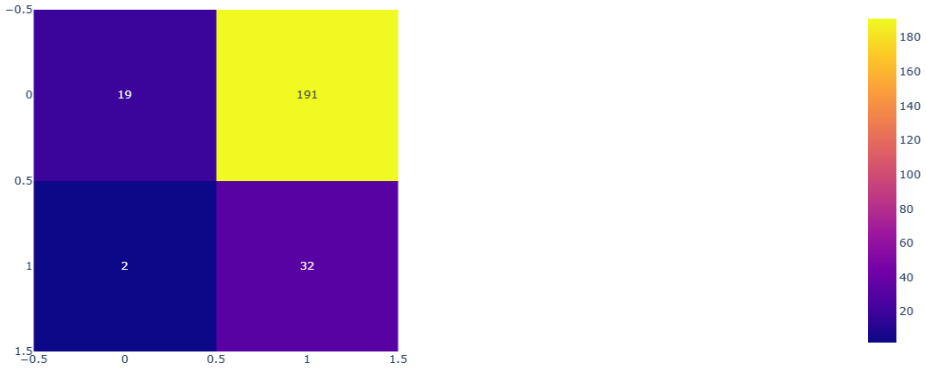


Tabla 6

Métricas de evaluación del clasificador Naïve Bayes

N°	Métricas de evaluación del Algoritmo	NB
1	Accuracy	0.209016
2	Recuperación	0.941176
3	precision	0.143498
4	roc_auc_score	0.515826

El modelo árbol de decisión obtuvo un 80% en la precisión, pero no logro encontrar las instancias de una clase especifica, la calidad de precisión es muy baja y la calidad del modelo es de un 49%.

Figura 19

Matriz de confusión del modelo árbol de decisión

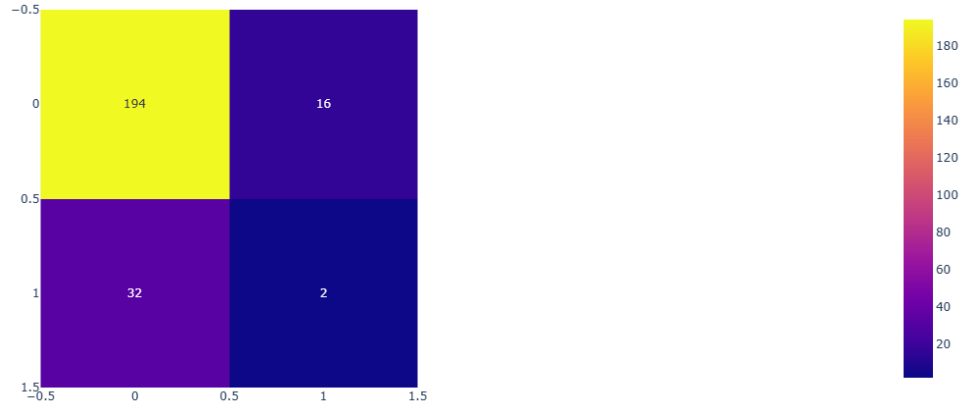


Tabla 7

Métricas de evaluación del clasificador árbol de decisión

	Métricas de evaluación del Algoritmo	AD
1	accuracy	0.803279
2	Recuperación	0.058824
3	precision	0.111111
4	roc_auc_score	0.491317

Es posible concluir que todos los clasificadores tuvieron un buen desempeño. Sin embargo, el algoritmo KNN se destaca entre los demás, ya que tiene los mejores resultados para todas las métricas de evaluación, a saber, 0.8197 de Accuracy, 0.22 de Precisión, 0.117 de Recall y un Característica Operativa del Receptor de 0.525 Por otro lado, es posible comprobar que no existen diferencias significativas en los valores obtenidos por los otros dos algoritmos, NB y AD.

Conclusiones

La importancia de la DM y sus técnicas para evaluar el rendimiento académico ha sido objeto de estudio en los últimos años. En este sentido, siguiendo las etapas de la metodología CRISP-DM se logró predecir el rendimiento académico en función de diversas dimensiones relacionadas, tales como la dimensión social y económica.

Es decir, se encontró la correlación entre el promedio semestral y las dimensiones sociales y económicos, los resultados obtenidos con este estudio se consideraron satisfactorios y muestran que el mejor modelo para la predicción del rendimiento académico de los estudiantes es el algoritmo K-vecinos más cercanos (KNN).

Asimismo, se aplicó la técnica Características categóricas de un codificador caliente para balancear las clases del atributo objetivo (promedio semestral).

El mejor resultado en la predicción lo obtuvo el algoritmo KNN con una Exactitud de 0.8197, una Precisión de 0.22, un Recall de 0.117647 y un Característica Operativa del Receptor de 0.525, por otro lado, el algoritmo Árbol de Decisión obtuvo una Exactitud de 0.803279, un recall 0.058824, una precisión 0.111111 y un Característica Operativa del Receptor de 0.491317, asimismo los atributos con una mayor correlación es la dependencia del estudiante y su situación socioeconómica.

Sugerencias

Con respecto al trabajo futuro, se recomienda recopilar más instancias con respecto a la Atributo de salida (Promedio semestral), en este caso, información de estudiantes, para mejorar la distribución de clases del conjunto de datos y, en consecuencia, eliminar la necesidad de utilizar técnicas de muestreo de datos. Además, se podrían aplicar otros algoritmos de Aprendizaje Profundo. Finalmente, se podrían experimentar más escenarios de prueba, utilizando diferentes técnicas de selección de características.

REFERENCIAS BIBLIOGRAFICAS

Abdallah, T., Al-Okaily, M., Alqudah, H., Matar, A., & Lutfi, A. (2020). Dataset on the Acceptance of e-learning System among Universities Students' under the COVID-19 Pandemic Conditions. *Data in Brief*, 106176. https://doi.org/10.1016/j.dib.2020.106176

Arias, F. G. (2012). *El proyecto de investigación introducción a la metodología científica* (Issue 1). https://doi.org/10.16309/j.cnki.issn.1007-1776.2003.03.004

Beyeler, M. (2018). *Machine Learning for OpenCV* (Issue june).

Bonaccorso, G. (2017). Machin Learning Algorithm. En *Biomass Chem Eng* (Vol. 49, Issues 23–6). www.packtpub.com

Bonaccorso, G. (2018). Mastering Machine Learning Algorithms. En *Psikologi Perkembangan* (Issue October 2013). https://doi.org/10.1017/CBO9781107415324.004

Chúmbez Rodríguez, M. F. (2017). Los estilos de aprendizaje y el rendimiento académico en la asignatura de lenguaje de los estudiantes del I Semestre – turno nocturno del Instituto de Educación Superior Tecnológico Público José Pardo" – La Victoria. *Universidad Nacional de Educación Enrique Guzmán y Valle*, 85.

Czibula, G., Mihai, A., & Crivei, L. M. (2019). S PRAR: A novel relational association rule mining classification model applied for academic performance prediction. *Procedia Computer Science*, *159*, 20–29. https://doi.org/10.1016/j.procs.2019.09.156

Dangeti, P. (2017). Statistics for Machine Learning: Techniques for exploring supervised, unsupervised, and reinforcement learning models with Python and R. En *Packt Publishing*.

de Pablos Pons, J. (2018). Las tecnologías digitales y su impacto en la Universidad. Las nuevas mediaciones. *RIED. Revista Iberoamericana de Educación a Distancia*, *21*(2), 83. https://doi.org/10.5944/ried.21.2.20733

del Cid, A., Méndez, R., & Sandoval, F. (2011). Investigación Fundamentos y Metodología. En *Prentice Hall*.

Feng, J. (2019). *Predecir el rendimiento académico de los estudiantes con el árbol de decisiones y la red neuronal*. 2004–2019.

Fernandes, E., Holanda, M., Victorino, M., Borges, V., Carvalho, R., & Erven, G. van. (2019). Educational data mining: Predictive analysis of academic performance of public school students in the capital of Brazil. *Journal of Business Research*, *94*(February), 335–343. https://doi.org/10.1016/j.jbusres.2018.02.012

Garbanzo Vargas, G. M. (2012). Factores asociados al rendimiento académico en estudiantes universitarios, una reflexión desde la calidad de la educación superior pública. *Revista Educación*, *31*(1), 43. https://doi.org/10.15517/revedu.v31i1.1252

Géron, A. (2017). Hands-on machine learning with Scikit-Learn and TensorFlow : concepts, tools, and techniques to build intelligent systems. En *O'Reilly Media*.

Giunchiglia, F., Zeni, M., Gobbi, E., Bignotti, E., & Bison, I. (2018). Mobile social media usage and academic performance. *Computers in Human Behavior*, *82*, 177–185. https://doi.org/10.1016/j.chb.2017.12.041

Gonzalez-Nucamendi, A., Noguez, J., Neri, L., Robledo-Rella, V., García-Castelán, R. M. G., & Escobar-Castillejos, D. (2021). The prediction of academic performance using engineering student's profiles. *Computers and Electrical Engineering*, *93*. https://doi.org/10.1016/j.compeleceng.2021.107288

Gron, A. (2019). *Hands-On Machine Learning with scikit learn keras&tensorflow (2nd ed.)*.

Hearty, J. (2016). *Advanced Machine Learning with Python*. www.packtpub.com

Holgado-Apaza, L. A. (2018). Detección de patrones de bajo rendimiento académico mediante técnicas de minería de datos de los estudiantes de la Universidad Nacional Amazónica De Madre De Dios 2018. En *Universidad Nacional del Altiplano* (Issue 051). http://repositorio.unap.edu.pe/handle/UNAP/9815

Imran, M., Latif, S., Mehmood, D., & Shah, M. S. (2019). Student academic performance prediction using supervised learning techniques. *International Journal of Emerging Technologies in Learning*, *14*(14), 92–104. https://doi.org/10.3991/ijet.v14i14.10310

Jansen, S. (2018). *Aprendizaje automático para el comercio algorítmico*.

Mcclure, N. (2017). *TensorFlow Machine Learning*. www.packtpub.com

Menacho Chiok, C. H. (2017). Predicción del rendimiento académico aplicando técnicas de minería de datos. *Anales Científicos*, *78*(1), 26. https://doi.org/10.21704/ac.v78i1.811

Méndez Aguirre, Ó. A., & Guillermo López, M. J. (2019). Técnicas de Machine Learning para la predicción de desempeño académico en el Desarrollo del espacio proyectivo del Pensamiento Espacial. *UNIVERSIDAD PEDAGÓGICA NACIONAL*, *23*(3), 2019. https://doi.org/10.1016/j.chb.2019.04.015

Naik, P., & Oza, K. (2019). *Python with Spyder* (Issue November).

Nelli, F. (2018). Análisis de datos de Python: Con Pandas, NumPy, y Matplotlib: Segunda edición. En *Python Data Analytics: With Pandas, NumPy, and Matplotlib: Second Edition*. https://doi.org/10.1007/978-1-4842-3913-1.

López Guevara, Ó. E., Raba Forero, J. E., & Turga Malagón, n. (06 de mayo de 2019). *Desarrollo E Implementaciónde Una Aplicación Web Y Móvilpara La Solicitud De Músicosen Bogotá - musicapp. revista ingeniería, matemáticas y ciencias de la información, 6*(11), 22. doi: http://dx.doi.org/10.21017/rimci.2019.v6.n11.a59

Orihuela Maita, G. Y. (2019). *Aplicación de Data Science para la Predicción del Rendimiento Académico de los Estudiantes de la Facultad de Ingeniería de Sistemas de la Universidad Nacional del Centro del Perú*. 114.

Pete, C., Julian, C., Randy, K., Thomas, K., Thomas, R., Colin, S., & Wirth, R. (2000). Crisp-Dm 1.0. *CRISP-DM Consortium*, 76.

Raschka, S., & Mirjalili, V. (2017). *Python Machine Learning, 2nd Ed.*

Raza Hasan, S. P. A. R. ez A. R. S. M. K. U. S. (2018). *Student Academic Performance Prediction by using Decision Tree Algorithm*. IEEE.

Roberto Hernandez Sampie. (2014). metodologia de la investigacion sexta edicion. En *sexta edicion: Vol. (5)2* (Issue 2).

Sarkar, D., Bali, R., & Sharma, T. (2018). Practical Machine Learning with Python A Problem-Solver's Guide to Building Real-World Intelligent

Systems. En *Practical Machine Learning with Rust*. https://doi.org/10.1007/978-1-4842-5121-8

Sebe, N., Cohen, I., Garg, A., & Huang, T. S. (2005). Machine Learning en la visión por computador. En *Machine Learning in Computer Vision*. https://doi.org/10.1007/1-4020-3275-7

Shukla, N. (2017). *Machine Learning con TensorFlow*. 240.

Singh, A., & Paul, S. (2020). *Hands-On Python Aprendizaje profundo para la Web*.

Tomasevic, N., Gvozdenovic, N., & Vranes, S. (2020). Una visión general y comparación de las técnicas de minería de datos supervisadas para la predicción del rendimiento de los exámenes de los estudiantes. *Computers and Education*, *143*(February 2019), 103676. https://doi.org/10.1016/j.compedu.2019.103676

Waheed, H., Hassan, S. U., Aljohani, N. R., Hardman, J., & Nawaz, R. (2020). Predecir el rendimiento académico de los estudiantes a partir de big data en VLE utilizando modelos de aprendizaje profundo. *Computers in Human Behavior*, *104*, 34. https://doi.org/10.1016/j.chb.2019.106189

Xu, X., Wang, J., Peng, H., & Wu, R. (2019). Predicción del rendimiento académico asociado a los comportamientos de uso de Internet usando algoritmos de aprendizaje automático. *Computers in Human Behavior*, *98*(January), 166–173. https://doi.org/10.1016/j.chb.2019.04.015

Yamao, E., Saavedra, L. C., Campos Pérez, R., & Huancas Hurtado, V. de J. (2018). *Predicción del rendimiento académico utilizando la minería de datos en estudiantes de primer año de la universidad peruana*. 151–160. https://doi.org/https://doi.org/10.24265/campus.2018.v23n26.05

ANEXOS

Anexo 1. Matriz de Consistencia

Título: "Predicción del rendimiento académico empleando algoritmos de aprendizaje supervisado en estudiantes del primer semestre de la carrera profesional de Ingeniería de Sistemas e Informática de la UNAMAD, 2020"				
PROBLEMAS	**OBJETIVOS**	**HIPÓTESIS**	**VARIABLES/ INDICADORES**	**METODOLOGIA**
General ¿Cómo se puede predecir el rendimiento académico de los estudiantes del primer semestre de la carrera profesional de Ingeniería de Sistemas e Informática de la UNAMAD mediante algoritmos de aprendizaje supervisado? **Específicos** 1. ¿Cuáles son los indicadores sociales, económicos y académicos con mayor incidencia para predecir el rendimiento académico de los estudiantes del primer semestre de la carrera profesional de	**General** Predecir el rendimiento académico mediante algoritmos de aprendizaje supervisado de los estudiantes del primer semestre de la carrera profesional de Ingeniería de Sistemas e Informática de la UNAMAD. **Específicos** 1. Determinar los indicadores sociales, económicos y académicos con mayor incidencia para predecir el rendimiento académico de los estudiantes del primer semestre de la carrera	**General** El rendimiento académico de los estudiantes del primer semestre de la carrera profesional de Ingeniería de Sistemas e Informática de la UNAMAD se puede predecir mediante algoritmos de aprendizaje supervisado. **Específicos** 1. El rendimiento académico de los estudiantes del primer semestre de la carrera profesional de Ingeniería de Sistemas e Informática de la UNAMAD se puede predecir mediante	**Variable independiente:** Algoritmos de aprendizaje supervisado **Dimensiones/ Indicadores:** D1. Mediciones de rendimiento - Matriz de confusión - Accuracy - Sensibilidad **Variable dependiente:** rendimiento académico **Dimensiones/ Indicadores:** D1. Sociales - Genero - Edad - Estado civil - Financiamiento de estudio	**Nivel:** Aplicativo **Enfoque:** Cuantitativo **Diseño:** No experimental de tipo correlacional transversal. **Población:** 861 registros históricos de estudiantes ingresantes a la carrera profesional de Ingeniería de Sistemas e Informática de UNAMAD recolectados desde el semestre académico 2010-1 al 2020-2 **Muestra:** muestra censal conformada por 861 registros

Ingeniería de Sistemas e Informática de la UNAMAD? 2. ¿Qué algoritmos de aprendizaje supervisado pueden predecir el rendimiento académico de los estudiantes del primer semestre de la carrera profesional de Ingeniería de Sistemas e Informática de la UNAMAD?	profesional de Ingeniería de Sistemas e Informática de la UNAMAD. 2. Determinar los algoritmos de aprendizaje supervisado que pueden predecir el rendimiento académico de los estudiantes del primer semestre de la carrera profesional de Ingeniería de Sistemas e Informática de la UNAMAD	indicadores sociales, económicos y académicos. 2. Los algoritmos de aprendizaje supervisado pueden predecir el rendimiento académico de los estudiantes del primer semestre de la carrera profesional de Ingeniería de Sistemas e Informática de la UNAMAD.	D2. Económicos - Contexto socioeconómico - Trabajo D3. Académicos - Modalidad de ingreso - Promedio semestral - Tipo de preparación - Bienestar psicológico	históricos de estudiantes ingresantes a la carrera profesional de Ingeniería de Sistemas e Informática de UNAMAD recolectados durante los semestres académicos 2010-1 al 2020-2

Anexo 2. Solicitud de datos de los estudiantes del 2020 y años anteriores

"Universidad Nacional Amazónica de Madre de Dios"
DIRECCIÓN UNIVERSITARIA DE ASUNTOS ACADÉMICOS
OFICINA DE REGISTROS ACADÉMICOS
"Año de la universalización de la salud"
"Madre de Dios Capital de la Biodiversidad del Perú"

INFORME N.º 009-2021-UNAMAD-DUAA-ORA/BRS

A : Dra. Lastenia Cutipa Chavez
DIRECTORA DE LA DIR. UNIV. DE ASUNTOS ACADÉMICOS
DE : Bach. Braulio Ramos Soncco
OFICINA DE REGISTROS ACADÉMICOS
ASUNTO : Solicito datos de estudiantes 2020 a anteriores años

Me dirijo a su despacho, con la finalidad de saludarla y remitirle informe sobre datos solicitados por el tesista Bach. Vargas Quispe, Alex Ali.

Que el área de asuntos académicos tiene los siguientes campos de datos género, edad, dirección, asistencia, modalidad de ingreso, N° de estudiantes por semestre desde el 2010-I hasta 2020-II y entre otros campos que requiere la investigación titulada "PREDICCION DE RENDIMIENTO ACADEMICO EMPLEANDO ALGORITMOS DE APRENDIZAJE SUPERVISADO EN ESTUDIANTES DEL PRIMER SEMESTRE DE LA CARRERA PROFESIONAL DE INGENIERIA DE SISTEMAS E INFORMATICA DE LA UNAMAD 2020"

Sin otro en particular es cuanto puedo informar para su conocimiento y fines pertinentes Atentamente;

Universidad Nacional Amazónica MDC
Vicerrectorado Académico
Dirección Universitaria de Asuntos Académicos
SECRETARÍA
Fecha: 25 ENE 2021
Hora: ..... Folio: 01
Reg: 227 Firma: ....

Bach. Braulio R amos Soncco
ESPECIALISTA ADMINISTRATIVO
OFICINA DE REGISTRO ACADÉMICOS

Anexo 3. Código de implementación de los algoritmos de aprendizaje supervisado

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import random
import plotly.express as px
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import MinMaxScaler
from sklearn.neighbors import KNeighborsClassifier
from sklearn.naive_bayes import GaussianNB
from sklearn.tree import DecisionTreeClassifier
from sklearn.neural_network import MLPClassifier
from sklearn.metrics import confusion_matrix
from sklearn.preprocessing import LabelBinarizer
from keras.models import Sequential

from  google.colab import drive
drive.mount('/content/drive')
datos= pd.read csv("/content/drive/MyDrive/prac ml/ExcelTeisV11.csv")
datos.head()
dummy Sex= pd.get dummies(datos["Sex"], prefix="Sex")
dummy_AdmisionType= pd.get_dummies(datos["AdmissionType"], prefix="AdmissionType")
dummy StudentDependency= pd.get dummies(datos["StudentDependency"], prefix="StudentDependenc
y")
dummy CivilStatus= pd.get dummies(datos["CivilStatus"], prefix="CivilStatus")
dummy_Preparation= pd.get_dummies(datos["UniversityPreparation"], prefix="UniversityPreparat
ion")
dummy Bienestar= pd.get dummies(datos["PsychologicalWell"], prefix="PsychologicalWell")
dummy trabajo= pd.get dummies(datos["StudentWorkCondition"], prefix="StudentWorkCondition")
dummy_Socioeconomico= pd.get_dummies(datos["Socioeconomic"], prefix="Socioeconomic")
info=datos.drop(["PsychologicalWell","Sex","Age","AdmissionType","StudentDependency","CivilS
tatus","UniversityPreparation","StudentWorkCondition","Socioeconomic","StudentWorkCondition"
],  axis=1)
info.head()
datos=pd.concat( [info,dummy Bienestar,dummy Sex,dummy AdmisionType,dummy StudentDependency,
dummy CivilStatus,dummy Preparation,dummy trabajo,dummy Socioeconomico],axis=1)
datos.head()
datos.shape
datos.groupby('WeightedAverageGrade').size()
X= datos.drop(['WeightedAverageGrade'],axis=1)
Y= datos['WeightedAverageGrade']
x_train, x_test, y_train, y_test=train_test_split(X,Y,random_state=5,train_size=0.7)
# K-Vécinos más cercanos
modelo_KNN= KNeighborsClassifier(n_neighbors=5)
modelo_KNN.fit(x_train,y_train)
Exactitud_KNN=  modelo_KNN.score(x_test,y_test)
Exactitud_KNN=round(Exactitud_KNN,4)*100
Exactitud_KNN
```

```
pred = modelo_KNN.predict(x_test)
MC= confusion_matrix(y_test,pred)
fig= px.imshow(MC, text_auto=True)
fig.show()
from sklearn.metrics import accuracy_score, confusion_matrix, recall_score, roc_auc_score, p
recision_score
pd.DataFrame(data=[accuracy_score(y_test, pred), recall_score(y_test, pred),
                   precision_score(y_test, pred), roc_auc_score(y_test, pred)],
                   index=["accuracy", "recall", "precision", "roc_auc_score"])

# Modelo: Naïve Bayes
modelo_NB = GaussianNB()
modelo_NB.fit(x_train,y_train)
Exactitud_NB= modelo_NB.score(x_test,y_test)
Exactitud_NB= round(Exactitud_NB,4)*100
Exactitud_NB
pred= modelo_NB.predict(x_test)
MC= confusion_matrix(y_test,pred)
fig= px.imshow(MC, text_auto=True)
fig.show()
from sklearn.metrics import accuracy_score, confusion_matrix, recall_score, roc_auc_score, p
recision_score
pd.DataFrame(data=[accuracy score(y test, pred), recall score(y test, pred),
                   precision score(y test, pred), roc auc score(y test, pred)],
                   index=["accuracy", "recall", "precision", "roc_auc_score"])

# Modelo Árbol de Decisión
modelo_AD= DecisionTreeClassifier(criterion='gini',max_depth=100)
modelo_AD.fit(x_train,y_train)

Exactitud_AD= modelo_AD.score(x_test,y_test)
Exactitud_AD= round(Exactitud_AD,4)*100
Exactitud_AD
pred= modelo_AD.predict(x_test)
MC= confusion_matrix(y_test,pred)
fig=px.imshow(MC,text_auto=True)
fig.show()
from sklearn.metrics import accuracy_score, confusion_matrix, recall_score, roc_auc_score, p
recision_score
pd.DataFrame(data=[accuracy_score(y_test, pred), recall_score(y_test, pred),
                   precision_score(y_test, pred), roc_auc_score(y_test, pred)],
                   index=["accuracy", "recall", "precision", "roc_auc_score"])
```